ACTA BIOMEDICA LOVANIENSIA

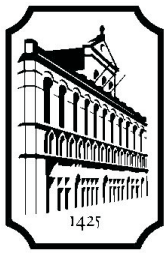
George KALEMA

FLEXIBLE REGRESSION MODELS FOR THE ANALYSIS OF HIERARCHICAL DATA FROM MEDICAL STUDIES

ACTA BIOMEDICA LOVANIENSIA 647
KU Leuven
Group Biomedical Sciences
Faculty of Medicine
Department of Public Health & Primary Care
L-BioStat

George KALEMA

FLEXIBLE REGRESSION MODELS FOR THE ANALYSIS OF HIERARCHICAL DATA FROM MEDICAL STUDIES



LEUVEN UNIVERSITY PRESS

Thesis submitted in partial fulfillment of the requirements for the degree of «Doctor of Biomedical Sciences (Biostatistics)»

Promoter:	Prof. dr. Geert Molenberghs
Co-promoters:	Prof. dr. Ziv Shkedy
	Prof. dr. Emmanuel Lesaffre
Chair:	Prof. dr. Bernadette Dierickx de Casterlé
Secretary:	Prof. dr. Francis Tuerlinckx
Jury members:	Prof. dr. Mia Hubert
	Prof. dr. Geert Verbeke
	Prof. dr. Philip Moons
	Prof. dr. Marc Aerts (UHasselt)
	Prof. dr. Michael Kenward (LSHTM, UK)

ISBN 978 94 6165 130 3
D/2014/1869/48
NUR: 876

"The heavens declare the glory of God; And the firmament shows His handiwork. Day unto day utters speech, And night unto night reveals knowledge. There is no speech nor language where their voice is not heard. Their line has gone out through all the earth, And their words to the end of the world."

Psalm 19:1-4 (KJV)

To my wife and children

Acknowledgement

A lot has prevailed during the preparation of this thesis. The following is my story. I arrived in Belgium, for the first time ever, on September 27, 2005, thanks to the VLIR scholarship I was granted to follow the Master of Science (MSc.) in Biostatistics programme for the academic year 2005-2007. I defended my MSc. thesis on November 6, 2007 after which I got the opportunity, starting from January 2008 to work at L-BioStat (then Biostatistical Center), in Leuven with dr. Kris Bogaerts and Prof. dr. Emmanuel Lesaffre, primarily on consultancy projects under a pre-doctoral programme. In June 2009, I officially started my PhD research under the supervision of Prof. dr. Geert Molenberghs, while continuing 40% of my time on consultancy with Kris. Let me pause right here and gratefully appreciate and acknowledge the VLIR-UOS scholarship programme and the tremendous lessons I have learned from Geert, Emmanuel and Kris over the years. Geert has especially ensured that this thesis is realized through (almost) weekly *touch-base* meetings together to a large extent with Samuel, Achmad and Mehreteab, and a never tiring effort, despite his busy schedule, to correct, guide and encourage me along the challenging path to attaining this PhD. Geert, I am sincerely grateful for the opportunity you accorded me to work with you and for your relentless efforts and expertise.

I have also been privileged to interact with Prof. dr. Geert Verbeke, director of L-BioStat, Kirsten Verhaegen, Ann Belmans, dr. Steffen Fieuws, dr.

Timothy Mutsvari, dr. Luwis Diya, dr. David Dejardin, dr. Samuel Iddi, dr. Mehreteab Aregay, dr. Chiara Forcheh, dr. Achmad Efendi, Anna Ivanova, Toon De Vis, Anikó Lovik and many others that have worked or are currently working at L-BioStat. Thanks to all these individuals, in their respective capacities, for every role they played.

Along the years, I met Jedidia Roose who became my wife and mother of our children (Elizabeth, Daniel and Aaron). My wife and children have been such a source of joy and strength for me while they endured so much during this PhD research. I gratefully thank them for their understanding of the season it has been and join them to celebrate the end of that season.

Lots of appreciations go to my jury members for their efforts, expertise and time sacrifice invested in reading, correcting and examining my thesis work.

I acknowledge the role played by my families in Belgium and Uganda. From the Ugandan side, my grandmother (Mrs. Tabitha Haumba), Aunt Loyce Munaba, mum (Ms. Elizabeth Nahidu), sis (Harriet Nakayenga) and all the others, have been instrumental in making this thesis a reality. Your continued support, encouragement, prayers and most importantly, the words of wisdom from hearts of experience have carried me through the years despite the spatial distance between us since I came to Belgium. I gladly submit this thesis to you as a testimony of your investment in my life over the years and a realization of the hope you had in me to rise, especially academically, to levels you were not able to. Ivan, my hope is that this thesis encourages you to work hard and keep the academic torch in our family blazing. Press through the tests, the testimony awaits. To my family (in-law) in Belgium, I very much appreciate your support and thank you, especially papa Patrick Roose and mama Vonnie Stringer, for being there for us through the good, the bad and the ugly.

I have also been blessed to know and meet many professing Christians in Belgium. I thank my brothers and sisters in Christ at Faith Worship Center in Alken, for an amazing period of growing and walking in the Lord, alongside the preparation of this thesis. I have always enjoyed fellowship (and swallowship) with brethren from Rhema Church in Sint-Truiden and Christ Centered Church in Leuven, among others. Individuals like Pastor Conley and Made-

laine, Jonathan and Sarah Weber, Pastor Seth Sakyi-Gyinae and wife, Jan van Deun, Seth Yeboah and wife, Linda and Tom Jackers, Chris and Christa Verniers, Charles and Doreen, and many more, have been such a blessing to my life and I am grateful.

Finally, many have been my weaknesses, mistakes and failures despite my desire to accomplish so much during the times of the PhD research. I presume to have learned a number of important lessons but more importantly, the grace of God has been so abundant through and through, and all I can say is, Ebenezer, so far has the Lord brought me.

George KALEMA
Leuven, June 2014

List of Papers

The contents of this thesis are mostly based on the following original publications;

Chapter 4: Kalema, G., Molenberghs, G., (2012). Pseudo-likelihood Methodology for Hierarchical Count Data. *Communications in Statistics, Theory and Methods. Accepted for Publication.*

Chapter 5: Kalema, G., Molenberghs, G., (2013). Second-order Generalized Estimating Equations for Correlated Count Data. *Submitted to Computational Statistics.*

Chapter 6: Kalema, G., Iddi, S., Molenberghs, G., (2013). The Combined Model: A Tool for Simulating Correlated Counts with Overdispersion. *Communication in Statistics (Computation and Simulation). Accepted for Publication.*

Chapter 7: Kalema, G., Molenberghs, G., (2013). Pairwise Likelihood as a Marginal Model Approach to Hierarchical Count Data using SAS Software. *Submitted to Journal of Statistical Software.*

Chapter 8: Kalema, G., Molenberghs, G., (2013). %GEE2Counts: A SAS Macro for Modeling Correlated Counts using Second-order Generalized Estimating Equations. *Submitted to Journal of Statistical Software.*

Chapter 9: Kalema, G., Molenberghs, G., (2013). The Combined Model as a Correlated or Overdispersed Count Data Simulator for Marginal Models; A SAS Implementation. *Journal of Statistical Software: Under review.*

Chapter 10: Kalema, G., Bogaerts, K., and Lesaffre, E., (2013). Estimating the Random Effects Distribution of Linear Mixed Models using SAS. *Submitted to Journal of Statistical Software.*

The author has also been involved in the following original publications;

Pittayapat, P., Galiti, D., Huang, Y., Dreesen, K., Schreurs, M., Couto Souza, P., Rubira-Bullen, I., Westphalen, F., Pauwels, R., **Kalema, G.**, Willems, G., Jacobs, R. (2012). An In-Vitro Comparison of Subjective Image Quality of Panoramic Views Acquired via 2D or 3D Imaging. *Clinical Oral Investigations*, art.nr. DOI 10.1007/s00784-012-0698-0.

Armstrong, P.W., Gershlick, A.H., Goldstein, P., Wilcox, R., Danays, T., Lambert, Y., Sulimov, V., Ortiz, F.R., Ostojic, M., Welsh, R.C., Carvalho, A.C., Nanas, J., Arntz, H.R., Halvorsen, S., Huber, K., Grajek, S., Fresco, C., Bluhmki, E., Regelin, A., Vandenberghe, K., Bogaerts, K., and Van de Werf, F., for the STREAM Investigative Team, (2013). Fibrinolysis or Primary PCI in ST-Segment Elevation Myocardial Infarction. *The new England Journal of Medicine*, **368**, 1379-87. DOI: 10.1056/NEJMoa1301092. *[I am listed in the Supplementary Appendix under Statistical Analysis Committee].*

Iddi, S., Molenberghs, G., Aregay, M., and **Kalema, G.** (2013). Empirical Bayes Estimates for Correlated Hierarchical Data With Overdispersion. *Pharmaceutical Statistics: Under review.*

Table of Contents

Abbreviations

Here, we give a list of the most often used abbreviations in the thesis.

AED	:	Anti-epileptic drug.
AIC	:	Akaike's information criterion.
ALR	:	Alternating logistic regression.
ARE	:	Asymptotic relative efficiency.
CM	:	Combined model.
EM	:	Expectation-maximization.
GEE	:	Generalized estimating equations.
GLM	:	Generalized linear model.
GLMM	:	Generalized linear mixed model.
i.i.d	:	Independent and identically distributed.
LMM	:	Linear mixed model.
MCAR	:	Mixing completely at random.
MC	:	Monte-Carlo.
MGamma	:	Multivariate Gamma.
ML	:	Maximum likelihood.
MMM	:	Marginalized multilevel model.
MP	:	Multivariate Poisson.
MSE	:	Mean squared error.
NB or NEGBIN	:	Negative-binomial.
NorTA	:	Normal to anything.

PGM	:	Penalized Gaussian mixture.
PL	:	Pseudo-likelihood.
RE	:	Random effect.
SAS	:	Statistical analysis system.
ZI	:	Zero-inflated.
ZINB	:	Zero-inflated negative binomial.
ZIP	:	Zero-inflated Poisson.

List of Tables

List of Figures

Part I

Introductory Material

Chapter 1

General introduction

1.1 Introduction

Research today results in enormous amounts of data being collected from research activity going on in so many fields of study, e.g., the airlines industry, medical research, banking, internet traffic, transportation, sports science, agriculture, environmental sciences, etc. Data, to a statistician, is characterized depending on the response/outcome variable of interest. For example, if the response is a time to an event of interest (e.g., time to relapse of a health condition), we refer to it as *survival data*. If binary (e.g., yes/no), the term *binary data* is used while *count data* refers to to data arising from a counting process in a given interval of time and therefore takes on non-negative integer values. Examples of count data may be, number of doctor visits, number of epileptic seizures, number of accidents, etc. These different characteristics can further be classified into two groups, namely, Gaussian (continuous) and non-Gaussian (binary, survival, count, etc.) outcomes.

Specific to medical research, on the one hand, a single outcome/response may be recorded for each study unit of interest (patient, subject, unit, etc) as well as the corresponding patient characteristics like gender, age, weight, height, etc. This is usually termed *univariate* or *cross-sectional data*. On the other hand, more than one observation may be recorded for each study unit.

This may be, for example, a study with many patients in which each patient is followed repeatedly over time resulting in the response(s) of interest being measured/recorded more than once. This alludes to the aspect of *repeated measures*. Also, if one response of interest is sequentially observed over time and data is recorded at specific time points during the study period, this is termed *longitudinal data*. An example of longitudinal data is when patients with epilepsy are followed over time and the number of their epileptic seizures recorded at certain time intervals. This results in data that exhibits correlation within a subject, meaning that measurements for a subject would be more related/similar than measurements between different subjects. When a set of measurements is collected from subjects that are structured in clusters, e.g., recording the weight of family (cluster) members across different families (clusters), this would be referred to as *clustered data*. Clustering arises due to such characteristics as, for example, familial relationships genetically which are shared between members of the same family. In analogy with longitudinal data, Spatial data (Cressie, 1991) arises when the time aspect is replaced by one or more spatial dimensions. For example, in agriculture, plots of land closer to each other may be more related, in terms of, e.g., soil fertility, erosion, etc. and therefore crop yield, than plots far away from each other. In another setting, several different variables/responses may be measured or observed from the same unit or subject, leading to *multivariate data*. *Time series data* differs from spatial data in that time series data has one natural direction of order (increase in time) while spatial data has two-dimensional or more directions and therefore no natural direction defines the ordering (Chernick, 2008). For all these different data structures, one thing is common, namely, that there are similarities/relationships between the units or subjects of interest that need to be considered. We therefore use the term *correlated data* generically as in Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005) to encompass these different structures.

In collecting the data, primary (observational, experimental, survey) or secondary (e.g., data from many publications into a single database) methods may be used. These data then have to be analyzed with statistical methods

to extract scientifically meaningful summaries or statements based on which valid conclusions and inferences can be made. Analysis of such data requires a good understanding of the mechanism that generated the data, or otherwise put, the design of the study. Understanding the design of the study is helpful as it contributes to determining the statistical methodology to use for the analysis. It is vital that the statistical method best reflects the design of the study and accounts for such intrinsic features as, correlation, overdispersion, underdispersion, excess zeros, etc., apparent in the data. These features are reflected upon in Chapter 3.

Depending on the response of interest (Gaussian or non-Gaussian), computation and analysis could be challenging in terms of time and the complexity of the models for correlated data. Given continuous correlated data, the well-known linear mixed model (LMM, Verbeke and Molenberghs, 2000) which assumes that the marginal distribution of the vector of responses for a subject, after conditioning on the normal distribution as the parent distribution for subject-specific effects, is a multivariate normal, is typically used for analysis. The marginal distribution is obtained by integrating out the random effects from the product of the distribution of the response given the random-effects distribution and the random-effects distribution. Although hierarchical in formulation, the straight forward marginalization of the LMM, thanks to the unique properties of the normal distribution, including that a product of two normal distributions is again a normal distribution, results in directly interpretable mean and covariance parameters. On the other hand, the case of non-Gaussian correlated data is unlike the Gaussian case due to the lack of a discrete distribution analogous to the normal distribution. This results in models that are more complex and prohibitive computationally especially with the increase in the number of replications per subject. This has been a point of focus over the years in research and has resulted in different methods which are either fully likelihood based (see, for example, Lipsitz *et al.*, 1991; Dale, 1986; Molenberghs and Lesaffre, 1994) or semi-parametric methods (e.g., Liang and Zeger, 1986; Geys *et al.*, 1998; McCullagh and Nelder, 1989). In the full likelihood approach, one specifies the joint distribution of

the measurements/outcomes for each cluster/subject. Models based on full likelihood methodology yield more efficiency than their semi-parametric counterparts but are computationally more involved, especially with large amounts of replications within a subject or unit.

Generally, when analyzing correlated data, 3 different modeling frameworks can be chosen from, depending on the objective of the study. In Section 5.3 of Molenberghs and Verbeke (2005), these frameworks or model families (marginal models, conditionally specified models, and subject-specific models) are presented and characterized. Other references in this light are Fahrmeir and Tutz (1994, 2001); Diggle *et al.* (2002), and, Lee and Nelder (2004). Briefly though, a marginal model is one where the marginal distribution of the response of interest is modeled as a function of covariates by conditioning the expectation of the response variable only on covariates. An example is a comparison of males and females in terms of the mean response or a contrast in the average number of epileptic seizures between patients who received a treatment versus patients in a control (placebo) group. One of the most used tools for the analysis of correlated non-Gaussian data, in the marginal model framework is generalized estimating equations (GEE, Liang and Zeger, 1986) while pseudo-likelihood (PL, Arnold and Strauss, 1991; Le Cessie and Van Houwelingen, 1994; Zhao and Joe, 2005; Molenberghs and Verbeke, 2005; Yi *et al.*, 2011) is a viable alternative. For count data, the negative binomial model (NB, Breslow, 1984; Lawless, 1987) is commonly used to account for overdispersion. Other examples are the zero-inflated Poisson (ZIP, Lambert, 1992) or zero-inflated negative binomial (ZINB, Ridout *et al.*, 2001) models for modeling excess zeros, or excess zeros and overdispersion, respectively. A random-effects or subject-specific model further conditions on unobserved or latent subject-specific random effects in addition to the covariates of interest. For a conditionally specified model, typically a transition model, the expectation of the response variable is modeled while conditioning on part or all of the remaining set of responses for a subject as well as covariates; in a transition model, conditioning is on past measurements.

A substantial amount of research has been done to account for correlation.

Breslow and Clayton (1993), and, Wolfinger and O'Connell (1993) extended the generalized linear modeling (GLM, McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972; Agresti, 2002) framework to the so-called generalized linear mixed model (GLMM) in which the correlation is accounted for by use of random effects. Again, unlike the continuous data case, the parameter estimates in the GLMM have a hierarchical interpretation and one can only derive marginal interpretations via additional calculations (see Section 19.4 of Molenberghs and Verbeke, 2005). Molenberghs *et al.* (2007, 2010) proposed a joint model for clustering and overdispersion through two separate sets of random effects. In the context of correlated count data, extensions of the univariate Poisson model to a multivariate version have also been proposed. Again, this has the advantage of a gain in efficiency as long as the model is correctly specified. However, use of a so-called Multivariate Poisson (MP) model is constrained by the complexity of the probability function to be calculated. This is because it involves extensive summations which may increase the computational burden with increase in the number of measurements per subject and/or sample size. Karlis (2003) uses the Expectation-Maximization (EM) algorithm to derive a MP distribution via a multivariate reduction technique. Karlis and Ntzoufras (2003) model sports data using a bivariate Poisson distribution. Kocherlakota and Kocherlakota (2001) apply a bivariate Poisson distribution to longitudinal data but with only two time points.

Indeed, the GLM, GLMM and the proposal by Molenberghs *et al.* (2007, 2010) are among many approaches taken by data analysts when interested in explaining the differences between outcome values based on measured or known patient characteristics (covariates), given univariate, correlated, or correlated and overdispersed data, respectively. One then uses a *regression model* to relate these covariates to the outcome variable and to answer different research objectives. One objective may be to identify the causal effect of one or a set of variables on the dependent/outcome variable. Another objective may be the prediction of a value of the response variable. Alternatively, regression may provide meaningful summaries between the response variable and the covariates. See Chapter 2 of Powers and Xie (2008) for more on regression

models. Univariate Gaussian outcomes are analyzed using linear regression models which extend to the LMM when data is correlated. Univariate non-Gaussian outcomes are analyzed using logistic regression, log-linear models, probit regression, etc., unified into the GLM framework and extend to the GEE or GLMM in case of correlation. The methods introduced above may fail in some specific circumstances. It is the goal of this thesis to address some of these issues and propose alternative tools that may be flexibly implemented due to the software that has also been provided. This thesis therefore makes the specific contributions listed in Section 1.2.

1.2 Thesis Contribution

While a lot of methodology is available to analyze data from clinical trials, epidemiological studies, as well as from other types of studies, with repeated or otherwise hierarchical data, most methods suffer from important drawbacks, including but not limited to: (a) not having a clinically relevant interpretation for the model parameters; (b) failure to account for correlation/association in some models; (c) computational complexity and numerical instability; (d) lack of modeling options, especially for count data; (e) failure to accommodate overdispersion in addition to correlation; (f) lack of methods to generate data from such models, which is highly relevant for Monte Carlo estimation and simulation techniques. The objective of this thesis is to alleviate several of these problems, by exploring pseudo-likelihood as an alternative to estimating equations, by extending generalized estimating equations, by coherently addressing the problem of simulating correlated Poisson data, and by providing software implementation that should be quite easy to use in SAS. More specifically, this thesis addresses the following specific objectives;

-1- *Propose a marginal model for correlated count data with valid inference not only in the marginal parameters but also in the association structure, using pseudo-likelihood methodology.*

In likelihood-based modeling frameworks, the marginal (log)likelihood is usually maximized to estimate the unknown parameters and make inference. In

analyzing correlated count data, one alternative is to assume the Multivariate Poisson (MP) distribution as the parent distribution for the vector of outcomes for each subject/cluster and then construct the likelihood. This would be the ideal situation as efficiency would be maximized. In practise, however, the MP distribution is constrained by the presence of summations in the probability function such that computational complexity grows with an increase in the dimension of the outcomes per subject and/or the sample size. Rather than specifying the full likelihood, the idea of pseudo-likelihood, or composite likelihood is to specify, for example, all univariate densities, or all pairwise densities over the set of all possible pairs within a sequence of repeated measures. To answer objective 1, pairwise likelihood is used so that we reduce the computational burden while capturing the pairwise associations.

-2- *To extend Generalized Estimating Equations (GEE) methodology to correlated count data by modeling both the mean and covariance simultaneously to permit inference on both the association structure and the marginal mean parameters.*

GEE is a very viable tool when one is not interested in the association, as long as the marginal mean is correctly specified. Consistent parameter estimates and standard errors are obtained even with the miss-specification of the working correlation assumption. Of course, severe miss-specification of the working assumption will compromise efficiency. Since it allows for the miss-specification of the working correlation structure, one cannot rely on the correlation estimates from GEE for formulating answers to scientific questions. By using the bivariate Poisson distribution, we develop estimating equations for correlated count data in which the mean and covariance are modeled simultaneously.

-3- *To extend the ideas put forward by Molenberghs* et al. *(2007, 2010) to a data simulation context in which correlated count data are generated with a pre-specified mean (possibly depending on covariates, such as treatment), and a pre-specified variance structure.*

Molenberghs *et al.* (2007, 2010) introduced the so-called combined model, combining the features of correlation/clustering and overdispersion. In the context of Poisson/count data, the GLMM or Poisson-normal model for correlated/-clustered data was combined with the Poisson-Gamma or Negative-binomial model for overdispersion to produce the Poisson-Gamma-Normal model. They also derived the closed form expressions of the GLMM and the combined model for the mean and variance given the normal and Gamma random effects. If one is interested in generating correlated data only, a comparison of the mean and variance derived by Molenberghs *et al.* (2007, 2010) from the GLMM with a desired marginal mean and variance can result in the possibility of generating hierarchical/correlated data. Should interest be in generating not only correlated but also overdispersed count data, the combined model can be used a generator by comparing the mean and variance derived by Molenberghs *et al.* (2007, 2010) given both the Gamma and normal random effects, and, the desired marginal mean and variance structure.

-4- *To provide the corresponding SAS software that implements the developments in 1, 2 and 3, respectively.*

In addition, we provide a SAS macro that can be used to estimate the random effects distribution on the linear mixed model based on Ghidey *et al.* (2004).

1.3 Outline of Thesis

This thesis is divided into four parts. The first part provides a general introduction to the subject matter of this thesis. Specifically, Chapter 1 has provided a general introduction and summarizes the contributions this thesis makes to the scientific community. Chapter 2 presents the datasets of motivating case studies while Chapter 3 provides a review of existing literature. The second part of this thesis presents this thesis' contribution in depth with Chapter 4 answering objective 1, Chapter 5 detailing objective 2, and Chapter 6 presenting the details in line with objective 3. Chapters 7, 8 and 9 make up the third part of this thesis and present the corresponding SAS macros that

implement the proposals in Chapters 4, 5 and 6, respectively. Additionally, Chapter 10 provides a SAS macro to estimate the random effects distribution of the linear mixed model following the method of Ghidey *et al.* (2004).

The thesis ends with Chapter 11 presenting some general conclusions on the thesis content, a discussion of some of the limitations of the proposed methods and the prospects for future research.

Chapter 2

Motivating Datasets

2.1 Introduction

We hereby present the datasets of motivating case studies that are analyzed in this thesis. In Section 2.2, data from a randomized double-blind, multicenter study on epilepsy is presented. The Jimma Infant Growth Study is introduced in Section 2.3 while another epilepsy study presented in Leppik *et al.* (1985) and Thall and Vail (1990) is described in Section 2.4. Section 2.5 introduces the whitefly study while Section 2.6 describes the Jimma Infant Growth Study but with a continuous response variable.

2.2 Count Data: Epilepsy Data

Data on epileptic seizures were obtained from a randomized double-blind, parallel group multicenter study to compare a placebo (treatment=0) and a new anti-epileptic drug (AED) in combination with one or two other AED's (treatment=1). The randomization of the epileptic patients took place after a 12-week stabilization period. The number of seizures were counted during this baseline period after which 45 patients were assigned to the placebo group and 44 to the AED group. Patients were then followed weekly for 16 weeks and then enrolled into a long-term open-extension study. Patient characteristics

including race, age (years), sex, height, and weight were also recorded. Some of the patients were followed for up to 27 weeks. The outcome of interest is the number of epileptic seizures experienced since the previous/last time (week) of recording. Molenberghs and Verbeke (2005), and references therein, give more detail and a report of earlier analyses of this set of data. The objective of the study was to assess if AED reduced the number of epileptic seizures over time relative to the placebo. Figure 2.1 shows the evolution of the number of seizures for each epileptic patient over the study period, while Figure 2.2 shows the distribution of the seizure counts over all study weeks and treatment groups.

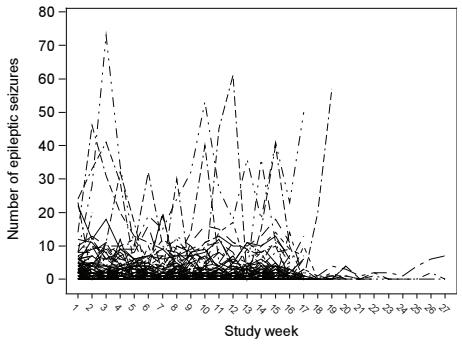


Figure 2.1: *Epilepsy data: Subject-specific profiles of the number of epileptic seizures over study weeks*

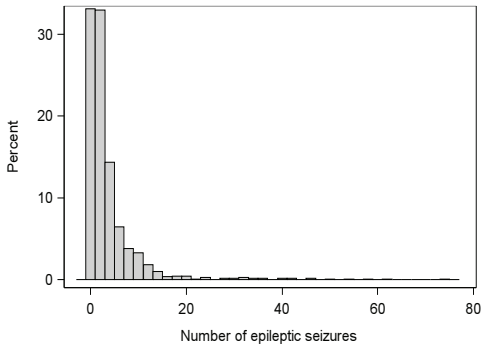


Figure 2.2: *Epilepsy data: Distribution of the number of epileptic seizures.*

Although not very easily observable, there generally seems to be a reduction in the number of seizures over time. There are also major differences in the seizure counts within patients but also between patients over time. We also observe from Figure 2.2 that the distribution of the seizure counts is quite skewed and that the majority of the counts were between 0 and about 15, although there was a count of up to 73 seizures in a week's time.

2.3 Count Data: Jimma Infant Growth Study

This dataset, also referred to as the Jimma Infant Survival Differential Longitudinal Growth Study, has been analyzed by Lesaffre *et al.* (1999) in the linear mixed models context, while Kassahun *et al.* (2012) has used it in the binary data framework in which they sought to identify risk factors for children being overweight, based on a dichotomization of the Body Mass Index (BMI). It is an Ethiopian study, set up to establish risk factors affecting infant survival and to investigate socio-economic, maternal, and infant-rearing factors that contribute most to the children's early survival. Children born in Jimma, Keffa and Illubabor, located in Southwestern Ethiopia were examined for their first year growth characteristics. At baseline (birth), there were a total of 7969 infants enrolled in the study, both singleton and twin live births inclusive. However, only singleton live births (7872 infants) at baseline are considered. The children were followed-up every two months, until the age of one year, thus $age = 0, 2, 4, 6, 8, 10$ and 12 months. Herein, we are interested in modeling the total number of days of diarrheal illness as a function of gender (1=Male, 0=Female), whether mother continued breastfeeding (1=Yes, 0=No) for the 12 months, whether mother sought medical help (1=Yes, 0=No), and place of residence (1=rural, 2=urban, 3=semi-urban). Figure 2.3 shows the evolution of the number of days of illness over the 12 months period for 399 (5%) randomly sampled infants while Figure 2.4 depicts the average number of days of illness over the 12 months by gender. From Figure 2.3, we observe a tendancy of the number of days of diarrheal illness to increase as the infant grows older. There is also a lot of variability observable within an infant, and likewise between infants as they evolve. Figure 2.4 further shows an increasing trend in the average number of days of diarrheal illness as the infants get older, with the females always having lower average counts than the males. Table 2.1 shows the number of infants whose responses were recorded over the 12 months period, by gender. As is typical of longitudinal studies, there is a reduction over time in the number of infants. As mentioned in Section 2.2, longitudinal studies very often have missing data. The Jimma study is no exception. Figure 2.5 shows some (20 out of 59) of the missingness patterns

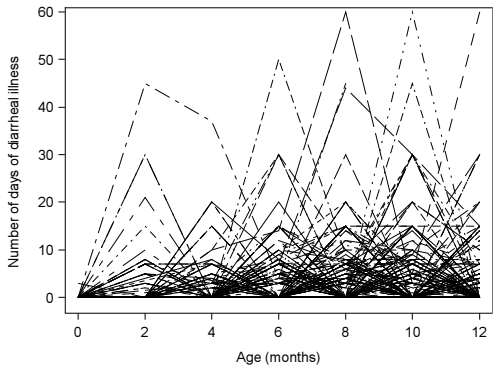


Figure 2.3: *Jimma Data: Infant-specific profiles of the number of days of diarrheal illness over age.*

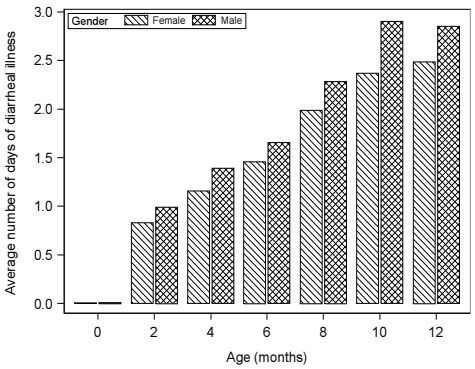


Figure 2.4: *Jimma Data: Average number of days of diarrheal illness by gender over age.*

Table 2.1: *Jimma data: Number of infants with observations by gender and age.*

	Age (months)						
Gender	0	2	4	6	8	10	12
Female	3865	3706	3570	3488	3401	3351	2920
Male	4007	3798	3656	3536	3455	3388	2972
Total	7872	7504	7226	7024	6856	6739	5892

present in the dataset. In general, both intermittent missingness and dropout as well as the first infant visit not having been at age=$0, 2, 4$ or 6 months are present. Since it is not our intention to deal with missing data in this thesis, we have assumed that the missingness mechanism is not related to the number of days of diarrheal illness observed and have excluded 525 infants with intermittent missingness.

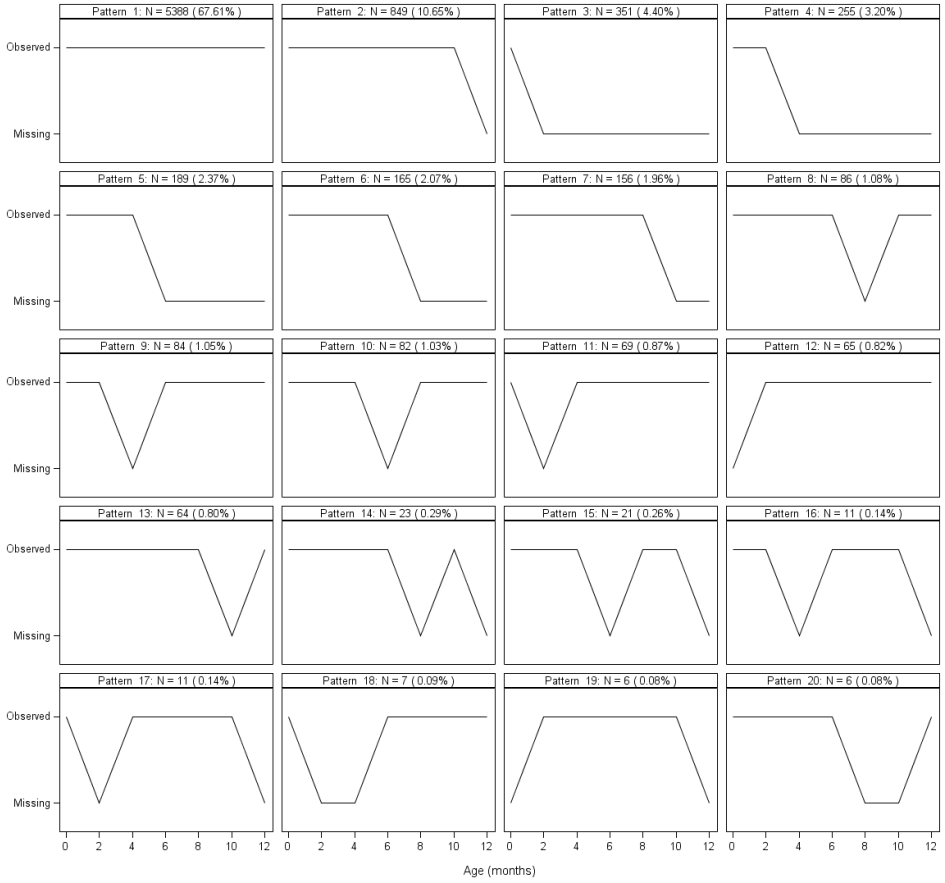


Figure 2.5: *20 of the 59 missing patterns in the Jimma dataset introduced in Section 2.3.*

2.4 Count Data: Epilepsy Data

This dataset is presented and analyzed by Leppik *et al.* (1985), and Thall and Vail (1990), among others. The data were obtained from a placebo-controlled clinical trial of 59 patients with epilepsy. These patients, suffering from simple or complex partial seizures, were enrolled in a randomized clinical trial that aimed at studying the effect of the anti-epileptic drug known as progabide on the number of epileptic seizures over time. In the study, 31 epileptic patients were randomized to the group that received progabide while 28 patients re-

ceived a placebo, as an adjuvant to the standard anti-epileptic chemotherapy. Progabide is an anti-epileptic drug whose primary mechanism of action is to enhance gamma-aminobutyric acid (GABA) content. GABA is the primary inhibitory neurotransmitter in the brain. Prior to receiving treatment, baseline data on the number of epileptic seizures during the preceding 8-week interval were recorded. Counts of epileptic seizures during 2-week intervals before each of four successive post-randomization clinic visits were recorded. The dataset also information on the patient identification, treatment (0=Placebo, 1=Progabide), age, baseline 8 week seizure count and the seizure count during the first, second, third and fourth 2-week time interval. Figure 2.6 shows the evolution of the number of seizures for each epileptic patient over the study period, while Figure 2.7 shows the distribution of the seizure counts over all week intervals and both treatment groups. The evolution of the average and median number of epileptic seizures between the consecutive two-weeks period by treatment are shown in Figures 2.8 and 2.9, respectively. There are differences in the seizure counts within patients but also between patients over time. Specifically, one patient seems to have an extreme number of seizure counts at all time points relative to the other profiles while another patient registered a rather distant number (76) of seizures at the third visit. We also observe from Figure 2.7 that the distribution of the seizure counts is quite skewed and that the majority of the counts were between 0 and about 20, although there was a count of up to 102 seizures in the first two-weeks (see Figure 2.6). From Figures 2.8 and 2.9, it can be seen that the progabide group has lower (mean or median) seizure counts except at the second two-weeks interval. The seizure counts seem to reduce, on average, over the study period for both treatment arms. It is common for longitudinal studies to have cases that at some point in the study drop out or miss some of the visits. For this dataset, however, all patients were observed at all the visits.

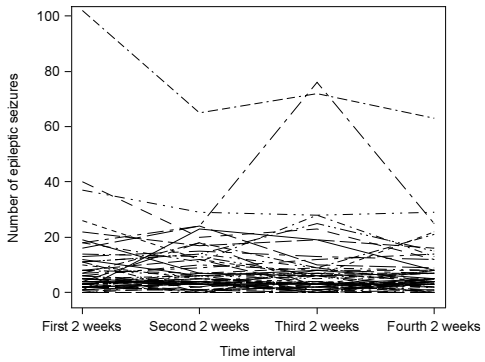


Figure 2.6: *Epilepsy Data: Subject-specific profiles of the number of epileptic seizures over study weeks.*

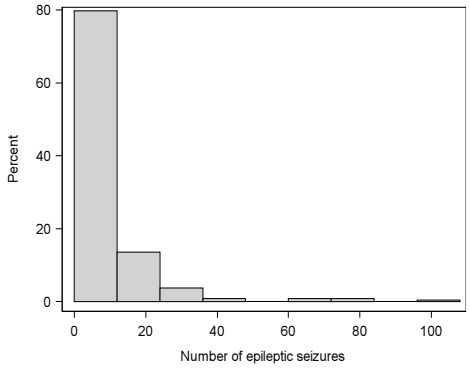


Figure 2.7: *Epilepsy Data: Distribution of the number of epileptic seizures.*

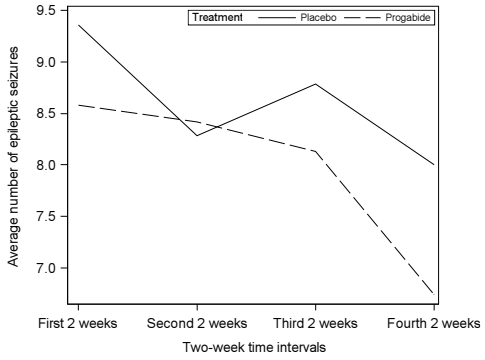


Figure 2.8: *Epilepsy Data: Average evolution of the number of epileptic seizures over study weeks by treatment.*

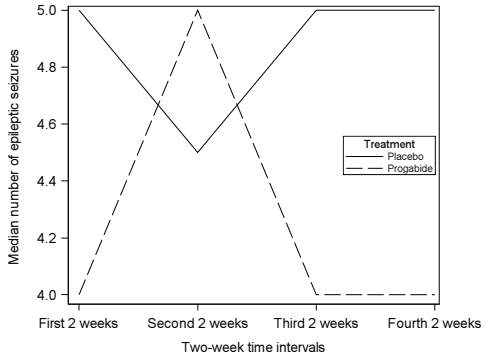


Figure 2.9: *Epilepsy Data: Median evolution of the number of epileptic seizures over study weeks by treatment.*

2.5 Count Data: The Whitefly Data

The whitefly dataset was reported in van Iersel *et al.* (2000) and also analyzed in Hall (2000) and Hall and Zhang (2004). It comes from a horticultural experiment that examined the effect of six methods (treatments) of applying the insecticide imidacloprid to poinsettia plants. Using a randomized complete block design, each treatment was applied to 18 experimental units that consisted of a trio of 18 poinsettia plants (54 plants in total). Data was then

collected repeatedly over a period of 12 weeks. The experimental units were randomly assigned to the 6 treatments in 3 complete blocks. Two outcomes, namely, the number of surviving adult whiteflies and the number of immature whiteflies were observed in the study. In this thesis, though, we focus on the latter, thus, the number of immature whiteflies after treatment out of a number of insects caged in one leaf per plant, prior to measurement of the response.

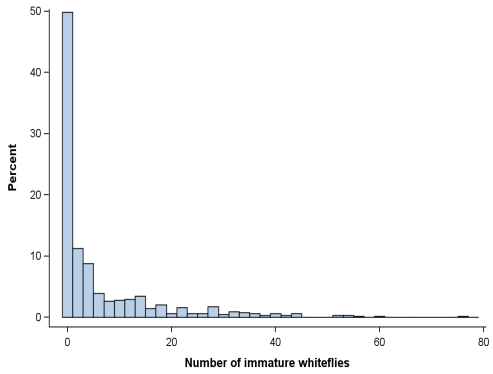


Figure 2.10: *Whitefly Data: Distribution of the number of immature whiteflies over all weeks and treatments.*

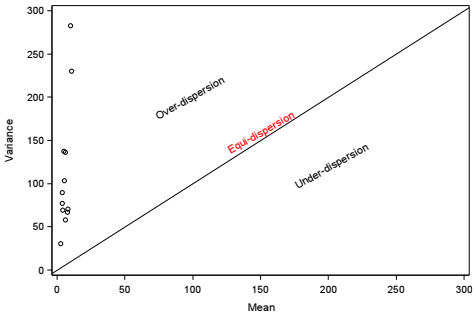


Figure 2.11: *Whitefly Data: Mean-variance relationship, over all treatments. Each dot represents a week of study.*

It can be seen from Figure 2.10 that almost 50% of the recorded observations (number of immature whiteflies) is zero, raising the issue of zero-inflated count data. It is also visible from Figure 2.11 that the variance expected from the Poisson distribution is greater than the mean, raising the issue of overdispersion. Another interesting feature of this dataset is the fact that observations were collected repeatedly over the 12 weeks alluding to the phenomenon of correlated count data.

2.6 Continuous Data: Jimma Infant Growth Study

The Jimma Infant Survival Differential Longitudinal Growth Study, introduced in Section 2.3, examined live births from 11 September, 1992 to 10 September, 1993 in one urban area and several rural areas in South-West

Ethiopia. It involved 8,000 households and the children were followed up approximately every two months starting immediately after birth over a study period of one year. As briefly mentioned in Section 2.3, it has been analyzed in different contexts. More specifically, Lesaffre *et al.* (1999) analyzed this data in the context the linear mixed model. They sought to identify the determinants of the growth of children in Ethiopia in terms of body weight (kg) as an indicator of health status. Herein, we follow along the lines of Lesaffre *et al.* (1999) and use a learning random sample of 495 children with 3070 observations to demonstrate how the random effects distribution of the linear mixed model can be estimated using a mixture of normal densities. This dataset is analyzed in Chapter 10 of the thesis.

Chapter 3

Literature Review

3.1 Introduction

In this chapter, a review of the existing methods for the analysis of both univariate and correlated data is presented. These methods were either used for the analysis of some of the datasets described in Chapter 2, provided foundations for extensions and/or have been implemented in SAS software, in this thesis. Section 3.2 describes the generalized linear modeling (GLM) framework, a class of fixed-effects models unifying linear, logistic and Poisson regression models, among others. GLM are usually used for the analysis of univariate or cross-sectional data whenever interest is in relating known covariates with a response variable. Overdispersion, a commonly encountered phenomenon in the analysis of non-normal data is also described in Section 3.2.1. Models for correlated data are discussed in Section 3.3. While a case for continuous longitudinal data is presented in Chapter 10, most of this thesis focuses on correlated count data.

3.2 The Generalized Linear Model

As briefly mentioned in Section 1.1, the GLM is a unifying (or generalizing) framework for several statistical models, e.g., linear, logistic and Poisson re-

gression models. This generalization is motivated by two reasons, namely, (a) cases that have range restrictions should not necessarily be assumed to be normally distributed; and (b) the mean, depending on the response variable in consideration, may not necessarily be taken as a linear combination of parameters, also due to range restrictions, but a certain function of the mean may be. Consider an example of count data that may be assumed to follow a Poisson distribution, with mean, say, λ. λ will not be expressed as a linear combination of covariates and unknown parameters (which would allow the mean to live on the whole real line $\mathbb{R}$) but $\log(\lambda)$ will be. Therefore, a GLM relates a function of the mean to the covariates and unknown parameters linearly, replacing data transformation which was a common approach pre-GLM (McCulloch and Searle, 2001). The GLM is commonly applied in the context of cross-sectional data and is available in many statistical software packages including, for example, R, SAS, SPSS, STATA, MATLAB, etc. This modeling framework is based on the so-called exponential family of distributions.

A random variable Y is said to belong to an exponential family of distributions (also known as exponential dispersion model by Jørgensen, 1987) if the density is of the form

$$f(y) \equiv f(y|\eta,\phi) = \exp\left\{\phi^{-1}[y\eta - \psi(\eta)] + c(y,\phi)\right\}, \tag{3.1}$$

for a specific set of unknown parameters η (natural parameter) and ϕ (dispersion parameter), and for known functions $\psi(\cdot)$ and $c(\cdot,\cdot)$. It is well known (Molenberghs and Verbeke, 2005), that the first two moments follow from the function $\psi(\cdot)$ as:

$$\mathsf{E}(Y) = \mu = \psi'(\eta), \tag{3.2}$$

$$\mathsf{Var}(Y) = \sigma^2 = \phi\psi''(\eta). \tag{3.3}$$

An important implication is that, in general, the mean and variance are related through $\sigma^2 = \phi\psi''[\psi'^{-1}(\mu)] = \phi\upsilon(\mu)$, with $\upsilon(\cdot)$ the so-called variance function, describing the mean-variance relationship.

Classical data types falling under the GLM framework are the normal or

Table 3.1: *Conventional exponential family members and extensions with conjugate random effects. An excerpt from Molenberghs* et al. *(2010).*

Element	notation	continuous	binary	count	time to event	
Standard univariate exponential family						
Model		normal	Bernoulli	Poisson	Exponential	Weibull
Model	$f(y)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$	$\pi^y\,(1-\pi)^{1-y}$	$\frac{e^{-\lambda}\lambda^y}{y!}$	$\varphi e^{-\varphi y}$	$\varphi\rho y^{\rho-1}e^{-\varphi y^\rho}$
Nat. param	η	μ	$\ln\left[\pi/(1-\pi)\right]$	$\ln\lambda$	$-\varphi$	
Mean function	$\psi(\eta)$	$\eta^2/2$	$\ln[1+\exp(\eta)]$	$\lambda=\exp(\eta)$	$-\ln(-\eta)$	
Norm. constant	$c(y,\phi)$	$\frac{\ln(2\pi\phi)}{2}-\frac{y^2}{2\phi}$	0	$-\ln y!$	0	
(Over)dispersion	ϕ	σ^2	1	1	1	
Mean	μ	μ	π	λ	φ^{-1}	$\varphi^{-1/\rho}\Gamma(\rho^{-1}+1)$
Variance	$\phi v(\mu)$	σ^2	$\pi(1-\pi)$	λ	φ^{-2}	$\varphi^{-2/\rho}\left[\Gamma(2\rho^{-1}+1)-\Gamma(\rho^{-1}+1)^2\right]$
Exponential family with conjugate random effects						
Model		normal-normal	Beta-binomial	Negative binomial	Exponential-gamma	Weibull-gamma
Hier. model	$f(y\|\theta)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\theta)^2}{2\sigma^2}}$	$\theta^y\,(1-\theta)^{1-y}$	$\frac{e^{-\theta}\theta^y}{y!}$	$\varphi\theta e^{-\varphi\theta y}$	$\varphi\theta\rho y^{\rho-1}e^{-\varphi\theta y^\rho}$
RE model	$f(\theta)$	$\frac{1}{\sqrt{d}\sqrt{2\pi}}e^{-\frac{(\theta-\mu)^2}{2d}}$	$\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$
Marg. model	$f(y)$	$\frac{1}{\sqrt{\sigma^2+d}\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2(\sigma^2+d)}}$	$(\alpha+\beta)\frac{\Gamma(\alpha)}{\Gamma(\alpha+y)}\frac{\Gamma(\beta)}{\Gamma(\beta+1-y)}$	$\frac{\Gamma(\alpha+y)}{y!\Gamma(\alpha)}\left(\frac{\beta}{\beta+1}\right)^y\left(\frac{1}{\beta+1}\right)^\alpha$	$\frac{\varphi\alpha\beta}{(1+\varphi\beta y)^{\alpha+1}}$	$\frac{\varphi\rho y^{\rho-1}\alpha\beta}{(1+\varphi\beta y^\rho)^{\alpha+1}}$
	$h(\theta)$	θ	$\ln[\theta/(1-\theta)]$	$\ln(\theta)$	$-\theta$	$-\theta$
	$g(\theta)$	$-\frac{1}{2}\theta^2$	$-\ln(1-\theta)$	θ	$-\ln(\theta)/\varphi$	$-\ln(\theta)/\varphi$
	ϕ	σ^2	1	1	$1/\varphi$	$1/\varphi$
	γ	$1/d$	$\alpha+\beta-2$	$1/\beta$	$\varphi(\alpha-1)$	$\varphi(\alpha-1)$
	ψ	μ	$\frac{\alpha-1}{\alpha+\beta-2}$	$\beta(\alpha-1)$	$[\beta\varphi(\alpha-1)]^{-1}$	$[\beta\varphi(\alpha-1)]^{-1}$
	$c(y,\phi)$	$-\frac{1}{2}\phi y^2-\frac{1}{2}\ln\left(\frac{2\pi}{\phi}\right)$	0	$-\ln(y!)$	$\ln(\varphi)$	$\ln\left(\varphi\rho y^{\rho-1}\right)$
	$c^*(\gamma,\psi)$	$-\frac{1}{2}\gamma\psi^2-\frac{1}{2}\ln\left(\frac{2\pi}{\gamma}\right)$	$-\ln\,B(\gamma\psi+1,\gamma-\psi\gamma+1)$	$(1+\gamma\psi)\ln\gamma-\ln\Gamma(1+\gamma\psi)$	$\frac{\gamma+\varphi}{\varphi}\ln(\gamma\psi)-\ln\Gamma\left(\frac{\gamma+\varphi}{\varphi}\right)$	$\frac{\gamma+\varphi}{\varphi}\ln(\gamma\psi)-\ln\Gamma\left(\frac{\gamma+\varphi}{\varphi}\right)$
Mean	$\mathsf{E}(Y)$	μ	$\frac{\alpha}{\alpha+\beta}$	$\alpha\beta$	$[\varphi(\alpha-1)\beta]^{-1}$	$\frac{\Gamma(\alpha-\rho^{-1})\Gamma(\rho^{-1}+1)}{(\varphi\beta)^{1/\rho}\Gamma(\alpha)}$
Variance	$\mathsf{Var}(Y)$	σ^2+d	$\frac{\alpha\beta}{(\alpha+\beta)^2}$	$\alpha\beta(\beta+1)$	$\alpha[\varphi^2(\alpha-1)^2(\alpha-2)\beta^2]^{-1}$	$\frac{1}{\rho(\varphi\beta)^{2/\rho}\Gamma(\alpha)}\left[2\Gamma(\alpha-2\rho^{-1})\Gamma(2\rho^{-1})-\frac{\Gamma(\alpha-\rho^{-1})^2\Gamma(\rho^{-1})^2}{\rho\Gamma(\alpha)}\right]$

continuous, binary, count, and time-to-event data. A summary of the relevant GLM quantities for these exponential family members is presented in Table 3.1, reprinted from Molenberghs *et al.* (2010). See same reference for a general discussion of these members. Typically, for one to make inference or answer a specific study question, either quasi-likelihood or full likelihood approaches are embarked on (McCullagh and Nelder, 1989; Molenberghs and Verbeke, 2005). The former restricts the model specification to the first two moments (Equations 3.2 and 3.3) and does not require the full distributional assumptions about the variance function while the latter is based on (3.1). In the full likelihood approach, the marginal (log)likeliood, expressed in terms of known covariates and unknown regression parameters, is usually maximized and the unknown regression parameters estimated using maximum likelihood estimation (MLE) methodology. We refer to McCulloch and Searle (2001), Molenberghs and Verbeke (2005), and McCullagh and Nelder (1989) for further discussion on estimation and inference for the GLM framework. Specifically, given a sample of K independent outcomes $Y_1, \ldots, Y_K$ together with corresponding p-dimensional vectors of covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K$, it is assumed that all Y_i are independent and identically distributed (i.i.d) with density $f(y_i|\eta_i, \phi)$, belonging to the exponential family, and with natural parameter η_i allowed to differ per observation i depending on covariates. Specification of the generalized linear model is completed by modeling the means μ_i as functions of the covariate values. More specifically, it is assumed that $\eta_i = h(\mu_i) = \boldsymbol{x_i}^\top \boldsymbol{\xi}$, for a known monotonic invertible link function $h(\cdot)$, and with $\boldsymbol{\xi}$ a p-dimensional vector of fixed unknown regression coefficients. The link functions $h(\cdot)$ may include, identity (commonly used for continuous or normal data) in which $\mu_i = \boldsymbol{x_i}^\top \boldsymbol{\xi}$, log (used for count data) in which $\log(\mu_i) = \boldsymbol{x_i}^\top \boldsymbol{\xi}$, inverse (used for time-to-event data) in which $-\mu_i^{-1} = \boldsymbol{x_i}^\top \boldsymbol{\xi}$ and the logit (used for categorical or multinomial, e.g., binary, binomial, data) in which $\log(\mu_i/(1-\mu_i)) = \boldsymbol{x_i}^\top \boldsymbol{\xi}$, among others. In practise, one uses linear regression for continuous data, logistic regression for e.g., binary or binomial data, Poisson regression for count data and Cox regression for time-to-event outcomes.

3.2.1 Modeling Overdispersion

Specific to the count case, it can be seen from Table 3.1 that the standard Poisson model implies that the mean and variance are equal (λ). In practice, though, it is not uncommon to find deviations from this implication by even just comparing the sample mean and sample variance. For example, it may be the case that the observed or sample variance is is greater than the observed mean, a phenomenon usually referred to as overdispersion. It is commonly encountered in data assumed to follow a binomial distribution, correlated or uncorrelated, correlated Bernoulli/binary random variables, correlated or independent observations arising from counting processes (Poisson data), and time-to-event/survival data. This is due to the mean-variance relationship inherent in the distributions that are assumed to be the data generating mechanisms. Overdispersion is, however, not an issue in the case of independent Bernoulli observations. Research has shown overdispersion to be caused by, for example, missing covariates and the presence of correlation between individual responses or clustering, among others. Depending on outcome type and model, not accounting for overdispersion may lead to bias in some or all parameters; it definitely biases precision estimates. The result is then usually smaller p-values for the statistical tests as well as, of course, confidence intervals that are narrower than should be if overdispersion were properly handled. This means that inference based on such statistical analyses is questionable and may be misleading. This indicates an inadequacy for the GLM to flexibly account for overdispersion and alternatives have been sought over the years. Hinde and Demétrio (1998a,b) study this phenomenon of overdispersion in general while Breslow (1984) and Lawless (1987) are specific to the Poisson case. Molenberghs and Verbeke (2005) present various model-based approaches that accommodate overdispersion, including the beta-binomial model (Skellam, 1948), the multivariate probit model (Dale, 1986; Molenberghs and Lesaffre, 1994), and certain versions of the generalized linear mixed model (Breslow and Clayton, 1993), to mention but a few. Overdispersion is usually accounted for by using the negative-binomial (NEGBIN) model. The NEGBIN model follows from a two-stage approach where in stage 1, we assume that

$Y_i|\zeta_i \sim \text{Poi}(\zeta_i)$ and in stage 2, that ζ_i is a random variable with $\mathsf{E}(\zeta_i) = \mu_i$ and $\mathsf{Var}(\zeta_i) = \sigma_i^2$. Then, it follows, using iterated expectations, that

$$\begin{aligned} \mathsf{E}(Y_i) &= \mathsf{E}[\mathsf{E}(Y_i|\zeta_i)] = \mathsf{E}(\zeta_i) = \mu_i, \\ \mathsf{Var}(Y_i) &= \mathsf{E}[\mathsf{Var}(Y_i|\zeta_i)] + \mathsf{Var}[\mathsf{E}(Y_i|\zeta_i)] = \mathsf{E}(\zeta_i) + \mathsf{Var}(\zeta_i) = \mu_i + \sigma_i^2. \end{aligned}$$

Similar exercise for, for example, the binary case reveals that purely bernoulli/binary data is unable to capture overdispersion while the Poisson distribution can clearly accomodate overdispersion should ζ_i be considered random. A common choice for the distribution of ζ_i is the gamma distribution. Combining the two stages and integrating (3.4) over the random effects ζ_i results in the NEGBIN as the marginal model. Fitting these models is done by maximizing the marginal likelihood

$$f(y_i) = \int f(y_i|\zeta_i) f(\zeta_i) d\zeta_i. \tag{3.4}$$

See Table 3.1 for a summary of the conditional, random effects and marginal distributions (3.4) resulting from the two-stage approach for some members of the exponential family. The corresponding means and variances of the marginal models are also given. It is not always the case that closed form expressions for the marginal models can be derived. However, in the case of the continuous, binary, count and time-to-event data, closed form expressions can be derived when so-called conjugate random effects (Molenberghs *et al.*, 2010) are assumed.

Although not as commonly encountered and studied as the case of overdispersion, another phenomenon called underdispersion, whereby the mean of the sample is greater than the variance, is apparent in literature (Ridout and Besbeas, 2004; Sellers and Shmueli, 2010; Cameron and Johansson, 1997; Castillo and Pérez-Casany, 1998).

3.3 Models for Correlated Data

In this section, a review of existing methods commonly used for the analysis of correlated data are presented. In Section 3.3.2, we briefly outline the classical linear mixed model (LMM) which is an extension of the linear regression model to correlated data. Ghidey *et al.* (2004) extend the LMM to the so-called penalized Gausssian mixture (PGM) LMM in order to flexibly estimate the random effects distribution of the LMM which is usually assumed to be normal. This is reviewed in Section 3.3.3 while an extension of the GLM to account for correlation is the topic of Section 3.3.4. Further, Sections 3.3.5 and 3.3.6 describe the so-called combined model (Molenberghs *et al.*, 2007, 2010), used to account for correlation and overdispersion simultaneously, and generalized estimating equations (Liang and Zeger, 1986; Ziegler, 2011; Winkelmann, 2008; Hardin and Hilbe, 2003), respectively. Finally, Section 3.3.7 is dedicated to pseudo-likelihood methodology, a viable alternative especially when the joint distribution of the response variable is cumbersome to evaluate.

3.3.1 Notation

Henceforth, the term *subject* will be used loosely to mean the independently replicated entity within which the repetition occurs; for example, patient, subject, cluster, or unit. We use the random variable Y_{ij} to denote the j-th observation of subject i's response variable, $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$. K therefore denotes the total number of subjects or sample size. Because the responses for each subject i are repeatedly recorded, subject i has an $n_i \times 1$ vector $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^\top$ of measurements. Further, let X_{ij} denote a $p \times 1$ vector of covariates, thus $\boldsymbol{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})^\top$, that are to be investigated for possible association with the response variable Y_{ij}. In matrix

notation,

$$\boldsymbol{X}_i = \begin{pmatrix} \mathbf{X}_{i1}^\top \\ \mathbf{X}_{i2}^\top \\ \vdots \\ \mathbf{X}_{in_i}^\top \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \dots & X_{i1p} \\ X_{i21} & X_{i22} & \dots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \dots & X_{in_ip} \end{pmatrix},$$

an $n_i \times p$ dimensional design matrix. It is important to mention that the covariates contained in $\boldsymbol{X}_i$ may be either changing over time j in which case they would be referred to as time-varying covariates, or otherwise time-stationary. In Section 5.1.1, attention is given to both time-stationary and time-varying covariates.

3.3.2 The Classical Linear Mixed Model

The classical linear mixed model (Harville, 1977; Laird and Ware, 1982) is defined as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\xi} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, i=1,\dots,K \tag{3.5}$$

where $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ are as defined in Section 3.3.1, and $\boldsymbol{Z}_i$ is $n_i \times q$ design matrix. $\boldsymbol{X}_i$ corresponds to the p-dimensional fixed effects vector $\boldsymbol{\xi}$ of unknown regression coefficients and $\boldsymbol{Z}_i$ to the q-dimensional random effects vector $\boldsymbol{b}_i$ of regression coefficients specific to subject i. $\boldsymbol{\epsilon}_i$ is a vector of residual components ε_{ij}. Model 3.5 contains both fixed and random effects, the former being population-averaged parameters while the latter pertain to subject-specific characteristics.

It is often assumed that $\boldsymbol{b}_i$ follows a q-dimensional normal distribution with mean vector zero and covariance matrix D. It follows from (3.5) that conditional on the random effects $\boldsymbol{b}_i$, $\boldsymbol{Y}_i$ is normally distributed with mean vector $\boldsymbol{X}_i\boldsymbol{\xi} + \boldsymbol{Z}_i\boldsymbol{b}_i$ and covariance matrix $\sigma^2\boldsymbol{I}_{n_i}$ thus $\boldsymbol{Y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta} \sim N(\boldsymbol{X}_i\boldsymbol{\xi} + \boldsymbol{Z}_i\boldsymbol{b}_i, \sigma^2\boldsymbol{I}_{n_i})$. Let $f(\boldsymbol{Y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})$ be the density function of $\boldsymbol{Y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}$ and $f(\boldsymbol{b}_i)$ be the density function corresponding to the $\boldsymbol{b}_i$, then the marginal density function of $\boldsymbol{Y}_i$ is obtained by integrating out the random effects from the joint

distribution as

$$f(\boldsymbol{Y}_i \mid \boldsymbol{\theta}) = \int f(\boldsymbol{Y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{b}_i) \mathrm{d}\boldsymbol{b}_i. \tag{3.6}$$

Inference is then based on estimators obtained by maximizing the marginal likelihood function

$$L_{ML}(\boldsymbol{\theta}) = \prod_{i=1}^{K} \left\{ \begin{array}{l} (2\pi)^{-n_i/2} \mid \mathbf{V}_i \mid^{-1/2} \\ \exp\left[-\frac{1}{2} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\xi})^\top \mathbf{V}_i{}^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\xi}) \right] \end{array} \right\}, \tag{3.7}$$

where $\boldsymbol{\theta}$ contains all the parameters $(\boldsymbol{\xi}^\top, \boldsymbol{\Lambda}^\top)^\top$, $\boldsymbol{\Lambda}$ being a vector of all variance and covariance parameters in $\mathbf{V}_i (= \boldsymbol{Z}_i D \boldsymbol{Z}_i^\top + \sigma^2 \boldsymbol{I}_{n_i})$ to be estimated. While inference on the fixed effects is robust to a departure from the common assumption of the normal distribution for the random effects (Butler and Louis, 1992; Verbeke and Lesaffre, 1996), this deviation may impact the efficient estimation of the fixed effects parameters as well as the standard errors based on the model specified and also influences inference on the random effects distribution. Verbeke and Lesaffre (1996) show that checking the validity of the random effects distribution assumption is hampered by the fact that the empirical Bayes estimates suffer from what they termed as shrinkage. A detailed explanation of the classical linear mixed model can be found in Verbeke and Molenberghs (2000).

3.3.3 The Penalized Gaussian Mixture Linear Mixed Model

Based on the argument that the random effects distribution may be too restrictive in practice, Ghidey *et al.* (2004) proposed the penalized gaussian mixture linear mixed model (PGMLMM) in which a flexible estimate of the random effects density is given. In the PGM linear mixed model, the distribution of the $\boldsymbol{b}_i$ in (3.5) is allowed to deviate from the normal distribution. Let $\boldsymbol{b}_i = \boldsymbol{R}\mathbf{s}_i$, $i = 1, \ldots, K$ where $\mathbf{s}_i$ are standardized random effects and $\boldsymbol{R}$ is a lower triangular matrix such that $\boldsymbol{R}\boldsymbol{R}^\top = \mathbf{D}$, the covariance matrix of the random effects. Then, assuming independent error terms,

$$f(\boldsymbol{Y}_i \mid \boldsymbol{b}_i = \boldsymbol{R}\mathbf{s}_i, \boldsymbol{\theta}) = N(\boldsymbol{X}_i \boldsymbol{\xi} + \boldsymbol{Z}_i \boldsymbol{R}\mathbf{s}_i, \sigma^2 \boldsymbol{I}_{n_i}). \tag{3.8}$$

Assume that the $\mathbf{s}_i$ extend over the interval $[-m, m]$ for the random intercept model and $[-m, m] \times [-m, m]$ for the random intercept and slope model but vanish outside this interval. To fit, for example, a random intercept and slope model, take a grid of equally spaced points on the interval $[-m, m]$ in both dimensions to construct the random effects distribution. Let these grids be the means of the basis Gaussian densities say $\mu_{1\jmath}, \jmath = 1, \ldots, J$ for the first dimension and $\mu_{2l}, l = 1, \ldots, L$ for the second dimension. Each basis Gaussian density will then have a mean $\boldsymbol{R}\boldsymbol{\mu}_{\jmath l}$ and a covariance matrix $\boldsymbol{R}\mathbf{D}_s\boldsymbol{R}^\top$ on the original scale of the random effects where $\boldsymbol{\mu}_{\jmath l} = (\mu_{1\jmath}, \mu_{2l})^\top$ and $\mathbf{D}_s = \text{diag}(\tau_1^2, \tau_2^2)$ is the mean and covariance matrix of the standardized $\mathbf{s}_i$, respectively. A key assumption is that the density of the random effects $\mathbf{b}$ can be well approximated by a mixture of Gaussian densities defined on the grid as

$$f(\mathbf{b} \mid \boldsymbol{\theta}) = \sum_{\jmath=1}^{J} \sum_{l=1}^{L} c_{\jmath l} N(\boldsymbol{R}\boldsymbol{\mu}_{\jmath l}, \boldsymbol{R}\mathbf{D}_s\boldsymbol{R}^\top) \tag{3.9}$$

where

$$c_{\jmath l} = \frac{\exp(a_{\jmath l})}{\sum_{k=1}^{J} \sum_{m=1}^{L} \exp(a_{km})}$$

are transformed elements of a $J \times L$ matrix of coefficients with the properties $\sum_{\jmath=1}^{J} \sum_{l=1}^{L} c_{\jmath l} = 1$ and $c_{\jmath l} > 0$. $a_{\jmath l}$ is as defined in vector $\mathbf{a}$ in $\boldsymbol{\theta}$ below (3.11). The marginal density of $\boldsymbol{Y}_i$ then becomes

$$f(\boldsymbol{Y}_i \mid \boldsymbol{\theta}) = \int f(\boldsymbol{Y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{b}_i \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{b}_i = \sum_{\jmath=1}^{J} \sum_{l=1}^{L} c_{\jmath l} N(\boldsymbol{\upsilon}_{i,\jmath l}, \mathbf{V}_i) \tag{3.10}$$

where $\boldsymbol{\upsilon}_{i,\jmath l}$ and V_i are the mean and covariance of the $\jmath l-$th specific normal density component defined as $\boldsymbol{\upsilon}_{i,\jmath l} = \boldsymbol{X}_i\boldsymbol{\xi} + \boldsymbol{Z}_i\boldsymbol{R}\boldsymbol{\mu}_{\jmath l}$ and $\mathbf{V}_i = \boldsymbol{Z}_i\boldsymbol{R}\mathbf{D}_s\boldsymbol{R}^\top\boldsymbol{Z}_i^\top + \sigma^2\boldsymbol{I}_{n_i}$ for subject i, respectively. The marginal log-likelihood is then a sum of (3.10) over all the K subjects thus

$$\ell(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^{K} \log\{f(\boldsymbol{Y}_i \mid \boldsymbol{\theta})\} \text{ where } \mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_K^\top)^\top.$$

This log-likelihood is then penalized for overfitting due to using large J and L. Parameters in the model are then estimated using maximum likelihood, from the following penalized marginal likelihood;

$$\ell_p(\boldsymbol{\theta}; \mathbf{Y} \mid \boldsymbol{\lambda}) = \ell(\boldsymbol{\theta}; \mathbf{Y}) - \left[\frac{\lambda_1}{2} \sum_j \sum_l (\Delta_1^e a_{jl})^2 + \frac{\lambda_2}{2} \sum_j \sum_l (\Delta_2^e a_{jl})^2 \right] \quad (3.11)$$

where $\boldsymbol{\theta}$ is a vector containing all the parameters to be estimated including: the fixed effects ($\boldsymbol{\xi}$), a stacked vector of the unique elements of $\boldsymbol{R}(\boldsymbol{\sigma_R})$, the error standard deviation on the log scale ($\log(\sigma)$) and the vector of coefficients $\mathbf{a} = (a_{11}, a_{12}, \ldots, a_{jl}, \ldots, a_{JL})$ upon which the difference operator $\Delta_w^e, w = 1, 2$ of order e for the w-th dimension, is applied. $\boldsymbol{\lambda}$ is a vector of smoothing parameters λ_1 and λ_2 corresponding to the two dimensions. The coefficients c_{jl} in (3.9) are jointly estimated with the other model parameters by maximizing (3.11) resulting in the *penalized Gaussian mixture* linear mixed model. A conditional maximization procedure is used to avoid possible convergence problems due to differences in the scale of the parameters in $\boldsymbol{\theta}$. In step 1, (3.11) is maximized with respect to $\mathbf{a}$ conditioning on initial values for the other parameters in $\boldsymbol{\theta}$. In step 2, conditioning on the updated $\mathbf{a}$ from step 1, the other parameters in $\boldsymbol{\theta}$ are updated. Iteration is done between steps 1 and 2 until convergence.

Ghidey *et al.* (2004) evaluate the penalized likelihood in (3.11) at several combinations of λ_1 and λ_2. They select the optimal penalty coefficient $\boldsymbol{\lambda}$ as the one that minimizes Akaike Information Criterion (AIC). Thus, for a given $\boldsymbol{\lambda}$, AIC($\boldsymbol{\lambda}$)=$-2\ell(\boldsymbol{\theta}; \mathbf{Y}) + 2\dim(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ where $\dim(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ is the effective degrees of freedom defined according to Gray (1992) as

$$\dim(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) = \operatorname{trace}((\mathbf{C} \cdot \mathbf{H}^{-1} \cdot \mathbf{C}^\top)^{-1} \cdot \mathbf{C} \cdot \mathbf{H}^{-1} \cdot \mathbf{I} \cdot \mathbf{H}^{-1} \cdot \mathbf{C}^\top)$$

where $\mathbf{C}$ is a contrast matrix, $\mathbf{H}$ and $\mathbf{I}$ are the observed Fisher information matrices based on the penalized and unpenalized log-likelihoods, respectively. For a more detailed explanation of the PGM approach, we refer to Ghidey *et al.* (2004) and Ghidey (2005).

Much as the PGM linear mixed model seems appealing, its implementation in statistical software and therefore the application is limited. The method has been implemented in MATLAB but the software is not available publicly should one be interested in applying this method. In Chapter 10, we provide a SAS implementation of this approach and illustrate its functionality based on a real life dataset introduced in Section 2.6 as well as simulated data. Our SAS implementation makes it quite easy and user-friendly to apply the PGM model.

3.3.4 The Generalized Linear Mixed Model

In dealing with correlated count data, the GLMM (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005) is a commonly used tool for analysis. Assume that, conditional upon a q-dimensional random effects vector $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$, the outcomes Y_{ij} are independent with densities of the form

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\xi}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij}, \phi)\right\}, \tag{3.12}$$

with

$$h[\psi'(\lambda_{ij})] = h(\mu_{ij}) = h[\mathsf{E}(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\xi})] = \boldsymbol{X}_{ij}^{\top}\boldsymbol{\xi} + \boldsymbol{Z}_{ij}^{\top}\boldsymbol{b}_i \tag{3.13}$$

where $h(\cdot)$ is a known link function as defined in Section 3.2, $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ are p- and q-dimensional vectors of known covariate values, respectively, $\boldsymbol{\xi}$ is a p-dimensional vector of unknown fixed regression coefficients, and ϕ is a scale (overdispersion) parameter. The GLMM, therefore, modifies the linear predictor in generalized linear models (3.1) to include unknown subject-specific or random effects in addition to the fixed effects. In practice, these random effects are usually assumed to follow a normal distribution, mainly for convenience and software availability. However, they can, in principle, be assumed to follow a different distribution than the normal. Specific to count data, the

GLMM takes the following form:

$$\begin{aligned} Y_{ij}|\boldsymbol{b}_i &\sim \text{Poi}(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= \boldsymbol{X}_{ij}\boldsymbol{\xi} + \boldsymbol{Z}_{ij}^{\top}\boldsymbol{b}_i, \\ \boldsymbol{b}_i &\sim N(0, D), \end{aligned} \tag{3.14}$$

whereby the conditional distribution of the observations from a subject i given the random effects $\boldsymbol{b}_i$ is Poisson with a rate parameter λ_{ij} that is log-linearly related to covariates. As mentioned before, the classical approach to obtaining the unknown parameter estimates is by maximizing the marginal (log)likelihood derived from (3.4) or more generally, (3.6). However, closed-form expressions for these integrals do not exist in all cases but Molenberghs *et al.* (2007) and Molenberghs *et al.* (2010) derived the marginal mean and covariance for the Poisson GLMM case, among others, as

$$\mathsf{E}(Y_{ij}) = \mu_{ij} = \ln(\lambda_{ij}) = \boldsymbol{X}_{ij}^{\top}\boldsymbol{\xi} + 0.5\boldsymbol{Z}_{ij}^{\top}D\boldsymbol{Z}_{ij}, \tag{3.15a}$$

$$\mathsf{Var}(\boldsymbol{Y}_i) = M_i + M_i(e^{\boldsymbol{Z}_i D \boldsymbol{Z}_i^{\top}} - J_i)M_i, \tag{3.15b}$$

respectively, where J_i is a matrix of 1's and M_i is a diagonal matrix with entries μ_{ij}. Higher-order marginal moments and the marginal joint distribution are derived in closed form for the Poisson case also in (Molenberghs *et al.*, 2010).

3.3.5 The Combined Model

The combined model as introduced by Booth *et al.* (2003); Molenberghs *et al.* (2007, 2010) has been shown to be an appealing tool for modeling not only correlated or overdispersed data but also for data that exhibit both these features. Unlike techniques available in the literature prior to the combined model, which use a single random-effects vector to capture correlation and/or overdispersion, the combined model allows for the correlation and overdispersion features to be modeled simultaneously, by using two sets of random effects.

In the context of count data, for example, the combined model naturally reduces to the Poisson-normal model, an instance of the generalized linear mixed

model in the absence of overdispersion and it also reduces to the negative-binomial model in the absence of correlation. Here, a Poisson model is specified as the parent distribution of the data conditional on a normally distributed random effect at the subject or cluster level and/or a gamma distribution at observation level. The CM is expressed as

$$Y_{ij} \sim \text{Poi}(\lambda^*_{ij}), \tag{3.16a}$$

$$\lambda^*_{ij} = \theta_{ij}\lambda_{ij} = \theta_{ij}\exp(\boldsymbol{X}_{ij}^\top\boldsymbol{\xi} + \boldsymbol{Z}_{ij}^\top\boldsymbol{b}_i), \tag{3.16b}$$

$$\boldsymbol{\theta}_i \sim \text{Gamma}(\text{mean} = \mathbf{1}, \text{variance} = \Sigma_i), \tag{3.16c}$$

$$\boldsymbol{b}_i \sim N(0, D), \tag{3.16d}$$

where θ_{ij}, the entries in $\boldsymbol{\theta}_i$, are the overdispersion parameters introduced at observation level. If the θ_{ij}'s are assumed to be independent as is often done in practice, then the association is only induced by the $\boldsymbol{b}_i$ and the θ_{ij} would cover the overdispersion not accounted for by the normal random effects. As such, Σ_i is reduced to a diagonal matrix. On the other hand, the θ_{ij} can be correlated such that Σ_i can take on more general structures, which implies the use of some form of multivariate Gamma (MGamma) distribution. The marginal mean and the marginal variance-covariance matrix take the form:

$$E(Y_{ij}) = \mu_{ij} = \theta_{ij}\exp(\boldsymbol{X}_{ij}^\top\boldsymbol{\xi} + 0.5\boldsymbol{Z}_{ij}^\top D\boldsymbol{Z}_{ij}), \tag{3.17a}$$

$$\text{var}(\boldsymbol{Y}_i) = M_i + M_i(P_i - \boldsymbol{J}_i)M_i, \tag{3.17b}$$

where $M_i = \text{diag}(\boldsymbol{\mu}_i)$ and

$$P_i = e^{\left(0.5\boldsymbol{Z}_i D\boldsymbol{Z}_i^\top\right)}\left(\Sigma_i + J_i\right)e^{\left(0.5\boldsymbol{Z}_i D\boldsymbol{Z}_i^\top\right)}.$$

Here, J_i is a matrix of ones. Note that we make use of the fact that the gamma random effects have unit mean.

3.3.6 Generalized Estimating Equations

Herein, we shall generically refer to all estimating equations as GEE while GEE1 denotes the method put forward by Liang and Zeger (1986) in which the correlation structure is calculated using the method of moments. Furthermore, GEE1.5 denotes the extension of GEE1 by replacing the moment-based estimation of the working correlation parameters with a second set of estimating equations (Prentice, 1988; Kim and Shults, 2010; Lipsitz *et al.*, 1991, are some of the many references). In GEE1.5, the two sets of estimating equations for the marginal mean and correlation structure are assumed orthogonal or independent, which simplifies the computational burden that would be encountered otherwise. Because these methods aim at obtaining marginal mean parameters that are consistent and asymptotically normally distributed, they permit inferences on the marginal mean regression parameters and standard errors even when the correlation structure is not correctly specified. As a result, and akin to GEE1, no scientific inference can be made on the correlation structures in GEE1.5 given that these association structures are allowed to be misspecified. On the other hand, allowing the two sets of estimating equations to be correlated results in GEE2 (Liang *et al.*, 1992; Zhao and Prentice, 1990; Prentice and Zhao, 1991). This implies that the first and second moments are then fully modeled while making working assumptions about the third- and higher-order moments. We revisit GEE1, GEE1.5 and GEE2 below.

Following from the theory of generalized linear models (GLM, Agresti, 2002; Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), the first two moments derived (Molenberghs and Verbeke, 2005, ch. 3) from a distribution that belongs to the exponential family of distributions are the mean and variance, expressed as,

$$\mathsf{E}(Y_{ij} \mid \boldsymbol{X}_i) = \mu_{ij}, \tag{3.18a}$$

$$\mathsf{Var}(Y_{ij} \mid \boldsymbol{X}_i) = V_{ij} = \phi \upsilon(\mu_{ij}), \tag{3.18b}$$

respectively, where ϕ is a scale parameter for the variance and $\upsilon(\cdot)$ is the variance function, which describes the dependency of the variance on the mean.

As explained in Section 3.2, (3.18a) is related to covariates in $\boldsymbol{X}_{ij}$ via a known link function $h(\cdot)$ (for example, log link for counts/Poisson data, logit or probit link for binary/binomial data) as $h(\mu_{ij}) = \boldsymbol{X}_{ij}^{\top}\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a $p \times 1$ vector of unknown regression parameters. If we let $\text{Corr}(Y_{ij}, Y_{ik} \mid \boldsymbol{X}_i) = \rho_{ijk}$, then $\mathsf{Cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = V_i(\boldsymbol{\xi}, \phi, \boldsymbol{\alpha}) = \phi C_i(\boldsymbol{\xi})^{\frac{1}{2}} R_i(\boldsymbol{\alpha}) C_i(\boldsymbol{\xi})^{\frac{1}{2}}$, where R_i is a correlation matrix, $C_i = \text{diag}(\upsilon(\mu_{ij}))$ is a diagonal matrix of variances and $\boldsymbol{\alpha}$ is a vector of correlation parameters. Specific to count data, Y_{ij} is assumed to follow a Poisson distribution with mean μ_{ij}, thus $Y_{ij} \sim \text{Poisson}(\mu_{ij})$. The Poisson density can be expressed as belonging to the exponential family by letting $\log \mu_{ij}$ to be the natural parameter, $\phi = 1$ and $\upsilon(\mu_{ij}) = \mu_{ij}$, as shown in Table 3.1. The marginal mean (3.18a) is then modeled in terms of covariates as $\log(\mu_{ij}) = \boldsymbol{X}_{ij}^{\top}\boldsymbol{\xi}$, hence also referred to as a log-linear or Poisson regression model and $\mathsf{Var}(Y_{ij} \mid \boldsymbol{X}_{ij}) = \mu_{ij}$. This model, therefore, specifically implies that the mean is equal to the variance, a phenomenon usually termed equi-dispersion. In practice though, deviations from this implication are common leading to underdispersion or overdispersion as explained in Section 3.2.1. While it is not this goal of this thesis to fully address over(under)-dispersion, we refer interested readers to, for example, Molenberghs *et al.* (2007, 2010) and references therein for further detail.

Given $(\phi, \boldsymbol{\alpha})$, Liang and Zeger (1986) iteratively solve the generalized estimating equation (GEE1) given by

$$\sum_{i=1}^{K} U_i(\boldsymbol{\xi}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}}^{\top} V_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0 \tag{3.19}$$

to obtain the estimates for $\boldsymbol{\xi}$ $(\widehat{\boldsymbol{\xi}})$. The iterative algorithm is as follows:

1. Obtain the starting values of $\widehat{\boldsymbol{\xi}}$ from fitting a GLM (thus assuming independence).

2. Given $\widehat{\boldsymbol{\xi}}$ or $\widehat{\boldsymbol{\xi}}^{(l)}$, calculate $(\hat{\phi}, \widehat{\boldsymbol{\alpha}})$ and therefore $\hat{V}_i = \hat{\phi} C_i^{\frac{1}{2}} \left(\widehat{\boldsymbol{\xi}}\right) R_i \left(\widehat{\boldsymbol{\alpha}}\right) C_i^{\frac{1}{2}} \left(\widehat{\boldsymbol{\xi}}\right)$ using the method of moments (Molenberghs and Verbeke, 2005, ch. 8).

3. Given $\hat{\phi}, \widehat{\boldsymbol{\alpha}}$ and $\hat{V}_i$, update $\widehat{\boldsymbol{\xi}}$ by using Fisher's scoring algorithm:

$$\widehat{\boldsymbol{\xi}}^{(l+1)} = \widehat{\boldsymbol{\xi}}^{(l)} - \left[\sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}}\right)^\top V_i^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}}\right)\right]^{-1} \left[\sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}}\right)^\top V_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i\right)\right].$$

The solution is obtained by iterating between steps 2 and 3 above until convergence meaning that the change in the parameter estimates satisfies (e.g., is less than) a pre-specified criterion. Assuming the marginal mean ($\boldsymbol{\mu}_i$) is correctly specified, consistent and asymptotically normally distributed parameter estimates $\widehat{\boldsymbol{\xi}}$ with mean $\boldsymbol{\xi}$ and variance-covariance matrix

$$\mathsf{Var}(\widehat{\boldsymbol{\xi}}) = I_0^{-1} I_1 I_0^{-1}, \tag{3.20}$$

where

$$I_0 = \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}}\right)^\top V_i^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}}\right) \text{ and}$$
$$I_1 = \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}}\right)^\top V_i^{-1} \mathsf{Var}(\boldsymbol{Y}_i) V_i^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}}\right),$$

are obtained. The variance estimator in (3.20) is commonly referred to as the sandwich estimator and results in the so-called empirically corrected standard errors. The parameter estimates and standard errors are asymptotically correct whether or not the working correlation structure is correctly specified.

Much as GEE1 has been found appealing by many data analysts and researchers, it has quite a number of issues associated with it. It is not our intention to exhaustively list them herein but refer to, for example, Lee and Nelder (2004), Lindsey and Lambert (1998), Crowder (1995), Sun *et al.* (2009), Wang and Carey (2004), among others, for further discussion of these issues. Specifically, GEE1 allowing the misspecification of the working correlation structure, thereby rendering it a nuisance, implies that the response vector ($\boldsymbol{Y}_i$) is given an arbitrary distribution and hampers checking assumptions about the covariance structure (Lee and Nelder, 2004). Specifying a covariance structure based on a model straightforwardly allows for inference on this covariance structure.

Further and more importantly, although consistent parameter estimates and standard errors can be obtained even with a misspecification of the working correlation assumption, careful estimation of the covariance/correlation is needed since it may affect the iterative updating of $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$, leading to a breakdown of the iterative procedure (Sun *et al.*, 2009). As an alternative estimation approach to the method of moments used by Liang and Zeger (1986) for the correlation structure, Kim and Shults (2010) use a two-stage approach to estimate the regression parameters $\widehat{\boldsymbol{\xi}}$ and the correlation parameters $\widehat{\boldsymbol{\alpha}}$. At stage 1, they iterate between (3.19), with $V_i^{-1} = C_i^{-\frac{1}{2}}(\boldsymbol{\xi})R_i^{-1}(\boldsymbol{\alpha})C_i^{-\frac{1}{2}}(\boldsymbol{\xi})$, and the estimating equation for $\boldsymbol{\alpha}$, namely:

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \left\{ \sum_{i=1}^{K} \boldsymbol{Z}_i^{\top}(\boldsymbol{\xi}) R_i^{-1} \boldsymbol{Z}_i^{\top}(\boldsymbol{\xi}) \right\} = 0, \tag{3.22}$$

where $\boldsymbol{Z}_i(\boldsymbol{\xi}) = (z_{i1}, \ldots, z_{in_i})^{\top}$ are the j-th Pearson residuals for subject i given by $z_{ij} = (y_{ij} - \mu_{ij})/\sqrt{v(\mu_{ij})}$ and evaluated at the current $\widehat{\boldsymbol{\xi}}$, until convergence. At stage 2, they plug $\widehat{\boldsymbol{\alpha}}$ from stage 1 into

$$\sum_{i=1}^{K} \text{trace} \left\{ \frac{\partial R_i^{-1}(\delta)}{\partial \boldsymbol{\alpha}} R_i(\boldsymbol{\alpha}) \right\} \Big|_{\delta = \widehat{\boldsymbol{\alpha}}} = 0, \tag{3.23}$$

and update $\widehat{\boldsymbol{\alpha}}$. The final $\widehat{\boldsymbol{\xi}}$ is obtained by solving (3.19) at the final $\widehat{\boldsymbol{\alpha}}$ from (3.23).

Yet another alternative to the method of moments for the correlation parameters is the proposal by Prentice (1988) in which estimating equations (3.19) are simultaneously solved with those of pairwise correlations ($\boldsymbol{\alpha}$) given by;

$$\sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\zeta}_i}{\partial \boldsymbol{\alpha}} \right)^{\top} H_i^{-1} \left(W_i - \boldsymbol{\zeta}_i \right) = 0, \tag{3.24}$$

where $W_i = (z_{i1}z_{i2}, z_{i1}z_{i3}, \ldots, z_{i,n_i-1}z_{i,n_i}, z_{i1}^2, z_{i2}^2, \ldots, z_{in_i}^2)^{\top}$ contains the products of subject i's pairs and squares of Pearson residuals $z_{is}z_{it}$ where $1 \leq s < t \leq n_i$, $H_i = \mathsf{Var}(W_i)$ and $\boldsymbol{\zeta}_i = \mathsf{E}(W_i)$. It is common for binary responses that the last n_i components of W_i, i.e., the squared residuals, are left out

due to the mean-variance relationship. Calculating $\mathsf{Var}(W_i) = \mathsf{Var}(z_{is}z_{it}) = \mathsf{E}(\{z_{is}z_{it}\}^2) - \mathsf{E}(z_{is}z_{it})^2$ requires

$$\begin{aligned}\mathsf{E}(\{z_{is}z_{it}\}^2) = \mathsf{E}(Y_{is}^2Y_{it}^2) - 2\mu_{it}\mathsf{E}(Y_{is}^2Y_{it}) + \mu_{it}^2\mathsf{E}(Y_{is}^2) - \\ 2\mu_{is}\mathsf{E}(Y_{is}Y_{it}^2) + 4\mu_{is}\mu_{it}\mathsf{E}(Y_{is}Y_{it}) - 2\mu_{is}^2\mu_{it}^2 + \\ \mu_{is}^2\mathsf{E}(Y_{it}^2) - 3\mu_{is}^2\mu_{it}^2 + \mu_{is}^2\mathsf{E}(Y_{it}^2).\end{aligned} \tag{3.25}$$

For binary response data, for example, and unlike the counts case, (3.25) simplifies (as $Y_{ij}^2 = Y_{ij}$) such that

$$\mathsf{Var}(z_{is}z_{it}) = 1 + (1-2\mu_{is})(1-2\mu_{it})\left(v(\mu_{is})v(\mu_{it})\right)^{-1/2}\psi_{ist} - \psi_{ist}^2,$$

where $\psi_{ist} = \mathsf{E}(z_{is}z_{it})$ are entries in $\boldsymbol{\zeta}_i$. The binary responses case, thus, turns out to be special since $\boldsymbol{\zeta}_i$ and H_i are then fully determined by the mean and correlation models without necessitating additional assumptions about higher-order moments. Generally though, obtaining matrix H_i involves the third and fourth moments of $\boldsymbol{Y}_i$, which are usually assumed to be equal to zero. Alternatives to this assumption may be sought depending on the type of response variable under consideration. In the context of binary response data, for example, Diggle *et al.* (2002) suggest $H_i = \mathrm{diag}\left[\mathsf{Var}(z_{i1}z_{i2}), \ldots, \mathsf{Var}(z_{i,n_i-1}, z_{i,n_i})\right]$, which only depends on $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\xi}}$ while they propose the use of the identity matrix for count responses. By using the identity matrix for the counts, there is a loss of efficiency in estimating $\boldsymbol{\alpha}$. They, however, argue that this efficiency loss has very little impact, in practice, on the estimation of $\boldsymbol{\xi}$ and yet simplifies computation by avoiding the estimation of additional higher-order parameters. Note that while H_i is a working variance-covariance matrix (meaning that it contains working assumptions usually being that the third- and fourth-order correlations are equal to zero, matters not whether it is correctly specified or not, only aides the estimation of the regression parameters $\boldsymbol{\xi}$ and cannot be used for formal inferences) for W_i; V_i in (3.19), on the other hand, is not a working covariance matrix because the second moments are specified by (3.24).

Note that Prentice (1988) assumes independence between (3.19) and (3.24). Again, this assumption implies a loss of efficiency but is defendable because

the consistency and asymptotic normality of the marginal mean regression parameters is not hampered by the misspecification of the correlation structure. Also important to mention is that the sets of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ both come with precision estimates and formal inference can be made on these parameters as long as the equations can be believed to have been correctly specified (Molenberghs and Verbeke, 2005). It may be desirable, however, to relax the independence assumption between (3.19) and (3.24). This may be the case if interest lies in the efficient estimation of both $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$. One may then be interested in minimizing the loss of efficiency accruing to the orthogonality assumption in GEE1.5. This leads us to the so-called second-order GEE (GEE2). Zhao and Prentice (1990) proposed an alternative to GEE1 or GEE1.5 in terms of correlations while Liang *et al.* (1992) used odds ratios, with both proposals aimed at modeling mutivariate binary responses. Prentice and Zhao (1991) extended the equations of Zhao and Prentice (1990) to the general case of discrete or continuous response vectors. They combine the response vector $\boldsymbol{Y}_i$ and the pairwise crossproducts W_i into one outcome vector $\mathbf{T}_i^\top = (\boldsymbol{Y}_i^\top, W_i^\top)$ and obtain the unknown parameters by setting

$$\begin{aligned} U(\boldsymbol{\Theta}) &= \sum_{i=1}^{K} U_i(\boldsymbol{\Theta}) = \sum_{i=1}^{K} \mathbf{D}_i^\top(\boldsymbol{\Theta})\Sigma_i^{-1}(\boldsymbol{\Theta})\mathbf{f}_i(\boldsymbol{\Theta}) \\ &= \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\xi}} & \frac{\partial \boldsymbol{\zeta}_i}{\partial \boldsymbol{\xi}} \\ \mathbf{0} & \frac{\partial \boldsymbol{\zeta}_i}{\partial \boldsymbol{\alpha}} \end{pmatrix} \begin{pmatrix} \mathsf{Var}(\boldsymbol{Y}_i) & \mathsf{Cov}(\boldsymbol{Y}_i, W_i) \\ \mathsf{Cov}(W_i, \boldsymbol{Y}_i) & \mathsf{Var}(W_i) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{\mu}_i \\ W_i - \boldsymbol{\zeta}_i \end{pmatrix}, \end{aligned} \tag{3.26}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\xi}^\top, \boldsymbol{\alpha}^\top)^\top$, to zero and solving the equations. Since solving (3.26) is computationally unattractive, Prentice and Zhao (1991) suggested specifying working variance matrices in Σ_i such that the third and fourth moments are expressed as functions of $\boldsymbol{\mu}_i$ and $\boldsymbol{\zeta}_i$. We show in Section 5.1.1 how we incorporate the bivariate Poisson distribution, rather than the multivariate Poisson distribution, into score equation (3.19) to allow scientific inference on the covariance as well as the mean parameters while modeling the covariance of $\boldsymbol{Y}_i$ at pair level. Use of the multivariate Poisson distribution would result

in a full likelihood approach, which would maximize efficiency. However, it is hampered by the complexity of the probability function due to summations which may increase the computational burden with increase in the number of measurements per subject and/or sample size (Karlis, 2003), the very reason estimating equations are being sought after. By using the bivariate Poisson distribution, closed form expressions for the third and fourth moments in (3.25) are easily obtainable and one could go ahead with the suggestion of Prentice (1988) but for correlated count data.

3.3.7 Pseudo-likelihood

In likelihood-based modeling frameworks, the marginal (log)likelihood is usually maximized to estimate the unknown parameters. For continuous longitudinal data, the marginal distribution and therefore the marginal (log)likelihood involves a product of the normal distributions for the data and the random effects resulting in a normal distribution as the marginal distribution. This presents no computational challenges and has been widely implemented in statistical software packages like SAS. For non-normal data, on the other hand, specification of the full likelihood can be very prohibitive computationally when measurement sequences are of moderate to large length (Molenberghs and Verbeke, 2005). Rather than specifying the full likelihood which in some cases is numerically and computationally cumbersome, the idea of pseudo-likelihood, or composite likelihood (Arnold and Strauss, 1991; Le Cessie and Van Houwelingen, 1994; Geys *et al.*, 1999; Aerts *et al.*, 2002; Zhao and Joe, 2005; Molenberghs and Verbeke, 2005; Yi *et al.*, 2011) is to replace the full or joint density by a simpler function assembled from a suitable factor. This is done by specifying, for example, all univariate densities, or all pairwise densities over the set of all possible pairs within a sequence of repeated measures in place of the full likelihood. In the case of pairwise densities, the likelihood contribution $f(y_{i1}, \ldots, y_{in_i})$ of subject i to the full likelihood is substituted with a product of $f(y_{is}, y_{it})$. For example, when $n_i = 3$, $f(y_{i1}, y_{i2}, y_{i3})$ is replaced by $f(y_{i1}, y_{i2}) \times f(y_{i1}, y_{i3}) \times f(y_{i2}, y_{i3})$ and the corresponding log-likelihood $\log f(y_{i1}, y_{i2}, y_{i3})$ is replaced by $\log f(y_{i1}, y_{i2}) + \log f(y_{i1}, y_{i3}) + \log f(y_{i2}, y_{i3})$.

In the general case of n_i measurements per subject i, the contribution of subject i to the log pseudo-likelihood is $p\ell_i = \sum_{1\leq s<t\leq n_i} \log f(y_{is}, y_{it})$ and the marginal log-pseudo-likelihood is given by

$$p\ell(\boldsymbol{\lambda}|\mathbf{Y}) = \sum_{i=1}^{K} \sum_{s<t} \log f(y_{is}, y_{it}), \tag{3.27}$$

where $\boldsymbol{\lambda}$ contains the unknown parameters estimated by setting the first derivative of (3.27) equal to zero.

In general, let S be a set of all $2^n - 1$ vectors of length n, consisting solely of zeros and ones, with each vector having at least one non-zero entry. Denote the subvector of $\boldsymbol{y}_i$ corresponding to the components of s that are non-zero as $\boldsymbol{y}_i^{(s)}$ and the associated joint density as $f_s(\boldsymbol{y}_i^{(s)}; \boldsymbol{\lambda})$. A pseudo-likelihood function is then defined by choosing a set $\delta = \{\delta_s | s \in S\}$ of real numbers, with at least one non-zero component. The corresponding log pseudo-likelihood then becomes

$$p\ell = \sum_{i=1}^{N} \sum_{s\in S} \delta_s \log f_s(\boldsymbol{y}_i^{(s)}; \boldsymbol{\lambda}). \tag{3.28}$$

When the components in (3.28) result from a combination of marginal and conditional distributions of the original distribution, a valid pseudo-likelihood function results. Specifically, the classical log-likelihood function, e.g., (3.27), is found by setting $\delta_s = 1$ if s is the vector consisting solely of ones, and 0 otherwise. With correct model specification, consistent and normally distributed estimators are obtained (Molenberghs and Verbeke, 2005), the variance-covariance matrix calculated using a sandwich estimator similar to that of GEE in (3.20).

Regularity conditions have to be invoked to ensure that (3.27) can be maximized by solving the pseudo-likelihood (score) equations. These can be found, for example, in Molenberghs *et al.* (2011). Importantly, because the components in (3.27) are derived from marginalizing the original distribution, a valid pseudo-likelihood function results. Details can be found in Joe and Lee (2008), who use weighting for reasons of efficiency in pairwise likelihood. Let $\boldsymbol{\lambda}_0$ be the true parameter. Under the aforementioned regularity condi-

tions, maximizing (3.27) or more generally (3.28), produces a consistent and asymptotically normal estimator $\widetilde{\boldsymbol{\lambda}}_0$ so that $\sqrt{N}(\tilde{\boldsymbol{\lambda}}_N - \boldsymbol{\lambda}_0)$ converges in distribution to $N_p[\mathbf{0}, I_0(\boldsymbol{\lambda}_0)^{-1} I_1(\boldsymbol{\lambda}_0) I_0(\boldsymbol{\lambda}_0)^{-1}]$. The regularity conditions, as well as explicit forms for $I_0(\boldsymbol{\lambda}_0)$ and $I_1(\boldsymbol{\lambda}_0)$, are provided in Appendix A.1. We refer to Molenberghs and Verbeke (2005) for further detail on the topic of pseudo-likelihood.

Troxel *et al.* (1998) used the product of all univariate distributions as an approximation to the full-likelihood. This significantly reduces the computational burden encountered in the full-likelihood approach yet still results in asymptotically unbiased estimators of the regression parameters. However, specifying univariate distributions for longitudinal data is based upon the unrealistic working assumption of no dependence between the several responses within a subject and may lead to highly inefficiently estimated regression parameters (Parzen *et al.*, 2007). Specifying the bivariate distribution for all the pairs of the responses from each subject may be a better approach. This has been used by (Parzen *et al.*, 2007) for longitudinal binary data with non-ignorable non-monotone missingness. We develop this approach in Chapter 4 in the context of marginal models for hierarchical or correlated count data.

Part II

Methodological Contributions

Chapter 4

Pseudo-likelihood Methodology for Hierarchical Count Data

4.1 Introduction

Count data collected repeatedly over time for the same subject are commonly encountered in scientific research. When collected only once per subject or at one time point, one usually assumes the data to be generated from a univariate Poisson distribution. Contemporary studies frequently aim at describing the evolution of subjects over time or observing more than one response from a single subject. Assuming a univariate Poisson distribution as the parent distribution of such data would ignore correlation and lead to erroneous inferences.

A lot of research has been done to account for correlation in count data. Breslow and Clayton (1993), and Wolfinger and O'Connell (1993) extended the generalized linear modeling (GLM) framework to the so-called generalized linear mixed model (GLMM) in which the correlation is accounted for by use of random effects. Molenberghs *et al.* (2007, 2010) propose a joint model for clustering and over-dispersion through two separate sets of random effects.

Extensions of the univariate Poisson model to a multivariate version have also been proposed. This has the advantage of a gain in efficiency as long as

the model is correctly specified. However, use of a so-called Multivariate Poisson (MP) model is constrained by the complexity of the probability function to be calculated. This is because it involves summations which may increase the computational burden with increase in the number of measurements per subject and/or sample size. Karlis (2003) uses the Expectation-Maximization (EM) algorithm to derive a MP distribution via a multivariate reduction technique. Karlis and Ntzoufras (2003) model sports data using a bivariate Poisson distribution. Kocherlakota and Kocherlakota (2001) apply a bivariate Poisson distribution to longitudinal data but with only two time points. In this chapter, we propose a pseudo-likelihood, taking the form of pairwise likelihood, to drastically simplify computational burden while retaining sufficiently high statistical efficiency. For each pair, a bivariate Poisson distribution is specified hence capturing the association between the two measurements. We restrict attention to each subject having at least 2 measurements recorded. We compare our proposal to generalizing estimating equations (GEE1, Liang and Zeger, 1986) based on a simulation with varying sample sizes (K) and number of measurements per subject i (n_i). Our proposal allows for n_i to differ between subjects but we assign equal n_i to all subjects in the simulations. We quantify the behavior of the two methods in terms of mean square error (MSE), variance, and the absolute bias of the estimators. Two cases worth investigating are (a) when there is no association in the data and, (b) where there exists association or when data is collected repeatedly per subject. An overview of generalized estimating equations and the general idea of pseudo-likelihood are given in Sections 3.3.6 and 3.3.7, respectively. We present our PL approach to modeling count data in Section 4.2 while Section 4.3 outlines the set-up as well as the results of the simulation study. Also presented is an application of the proposal to a clinical trial study in epileptic seizures.

4.2 A Model for Hierarchical Count Data

Inference in a good number of longitudinal studies is primarily based on marginal parameters. Using classical maximum likelihood methodology then

necessitates the full specification of the joint distribution for $\boldsymbol{Y}_i$. In the context of discrete data, one needs to specify the first-order moments as well as all higher-order moments (Molenberghs and Verbeke, 2005) which often is computationally restrictive for high-dimensional vectors of correlated data. With primary interest placed on the marginal parameters, however, tools like GEE and pseudo-likelihood (PL, Arnold and Strauss, 1991; Le Cessie and Van Houwelingen, 1994; Zhao and Joe, 2005; Molenberghs and Verbeke, 2005; Yi *et al.*, 2011) have been proposed and implemented in statistical software. These two tools still allow for within-subject dependence but yet are computationally more practical relative to full likelihood.

Assume that there are K independent subjects in a study with subject i having a measurement Y_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$ and a corresponding $q \times p$ known design matrix $\mathbf{X}_i$. Denote the responses of subject i at any given pair of time points, s and t as Y_{is} and Y_{it}, respectively, $1 \leq s < t \leq n_i$. Further, assume that W_k are independent Poisson random variables with means θ_k, $k = s, t$ or st. The random variables $Y_{is} = (W_{is} + W_{ist})$ and $Y_{it} = (W_{it} + W_{ist})$ then follow a bivariate Poisson distribution, i.e., $(Y_{is}, Y_{it}) \sim BP(\theta_{is}, \theta_{it}, \theta_{ist})$ given by

$$f(y_{is}, y_{it}) = e^{-(\theta_{is}+\theta_{it}+\theta_{ist})} \frac{\theta_{is}^{y_{is}}}{y_{is}!} \frac{\theta_{it}^{y_{it}}}{y_{it}!} \sum_{k=0}^{\min(y_{is}, y_{it})} \binom{y_{is}}{k} \binom{y_{it}}{k} k! \left(\frac{\theta_{ist}}{\theta_{is}\theta_{it}}\right)^k. \quad (4.1)$$

Let $\theta_{is}^{\star} = \theta_{is} + \theta_{ist}$ and $\theta_{it}^{\star} = \theta_{it} + \theta_{ist}$ where $\log(\theta_{is}) = \boldsymbol{X}_{is}\boldsymbol{\xi}$ and $\log(\theta_{it}) = \boldsymbol{X}_{it}\boldsymbol{\xi}$. Marginally, $Y_{is} \sim \text{Poisson}(\theta_{is}^{\star})$, $Y_{it} \sim \text{Poisson}(\theta_{it}^{\star})$ and θ_{ist} is the covariance between subsequent pairs of the random variables Y_{is} and Y_{it}. The marginal log pseudo-likelihood takes the form (3.27). Estimation of the parameters in $\boldsymbol{\lambda} = (\boldsymbol{\xi}, \theta_{ist})^{\top}$ is done in SAS/IML® using the Newton-Raphson (NR) algorithm; a macro has been written to this effect. See Appendix A.2 for the gradient and Hessian functions of the log PL function in (3.27), with respect to the unknown parameters in $\boldsymbol{\lambda}$, which are supplied to the NR optimization step.

Note that we have formulated a bivariate model only, even though our aim is to analyze hierarchical data with more than two repetitions. Fortunately, the use of GEE and PL methodology obviates the need to explicitly specify the

higher-order joint distributions. Also note that we assume the covariance (θ_{ist}) to be the same for all subjects and pairs ($=\theta_{st}$). This bears resemblance to an exchangeable correlation structure, but one must remember that, because the variance depends on the mean, the corresponding correlations will fluctuate with the mean, even though the covariances may be constant. The exception is when the mean is constant as well, in which case a classical exchangeable correlation matrix will follow. This assumption of equal covariance term can however be relaxed.

4.3 Simulation Study

Simulations have been done to compare the performance of GEE1 and our proposed pseudo-likelihood approach in the cases of both correlated and independent outcomes. We study the effect of varying sample sizes and number of measurements per subject for GEE1 with an exchangeable working correlation structure in comparison to pseudo-likelihood, based on 1000 simulations. The absolute bias, MSE, and the percent samples for which convergence has been reached, quantify the behavior of the two methods.

4.3.1 Design of Simulation Study

4.3.1.1 Simulation of Independent Data

We generated data for $K = 10, 100, 1000, 10000$ subjects, assuming the following model:

$$\mu_{ij} = \exp(\xi_0 + \xi_1 * \text{trt}_i + \xi_2 * \text{time}_{ij} + \xi_3 * \text{trt}_i * \text{time}_{ij}), \tag{4.2a}$$

$$Y_{ij} \sim \text{Poisson}(\mu_{ij}), \tag{4.2b}$$

for subjects $i = 1, \ldots, K$ and measurements $j = 1, \ldots, n_i$. The subjects are equally distributed across the two treatment groups ($\text{trt}_i = 0$ or 1) and time_{ij} is the ordering of the j^{th} measurement within subject i. Further, n_i was fixed to values of $2, 4, 8$, and 16 for all subjects within a given run for simulation purposes, even though our method allows for varying cluster sizes.

The regression parameters were specified as $\xi_0 = 1.4531$, $\xi_1 = -0.1869$, $\xi_2 = -0.0328$, and $\xi_3 = 0.0195$.

4.3.1.2 Simulation of Dependent Data

To generate dependent data, a subject-specific intercept b_i is introduced to (4.2a), changing it to

$$\mu_{ij} = \exp\left(\xi_0 + b_i + \xi_1 * \text{trt}_i + \xi_2 * \text{time}_i + \xi_3 * \text{trt} * \text{time}_{ij}\right). \tag{4.3}$$

First, the fixed-effect parameters were specified as in the case of no association in Section 4.3.1.1, while b_i is a subject-specific parameter that was assumed to follow a normal distribution with zero mean and a variance of 0.25^2, thus $b_i \sim N(0, 0.25^2)$. Datasets of varying sample sizes and cluster sizes were then generated from (4.3) and the "true" marginal parameters obtained by fitting a univariate Poisson model ignoring the correlation. The parameters obtained are consistent though the efficiency with which they are estimated is compromised. Note that data are generated from a hierarchical model to which marginal models are then fitted. This implies that the true values for the ξ parameters in (4.3) do not correspond to the true values for the marginal model. To deal with this issue, very large sample sizes were generated (starting from 1000 and going up all the way to 250,000) using the hierarchical model and then the subsequent marginal model was fitted. For the largest sample sizes, very stable estimates were obtained. These values, $\xi_0^{(m)} = 1.5807$, $\xi_1^{(m)} = -0.1881$, $\xi_2^{(m)} = -0.0340$, and $\xi_3^{(m)} = 0.0192$ were used to calculate the bias in the case of dependent data. The superscript (m) refers to 'marginal.'

4.3.2 Results

A comparison between GEE1 and PL is done in the context of hierarchical count data. GEE1 has been widely implemented in statistical software like SAS and R. Our proposed PL approach is implemented in SAS and a macro is available from the authors' web pages.

Not surprisingly, for independent data, GEE1 and PL parameter estimates

are very similar (Table 4.1), with differences especially occurring in Table 4.1 for $K = 10$, $n_i = 2$. This is also evident in Figures 4.1 and 4.2 which present the MSE of both GEE and PL. PL, however, has the covariance parameter θ_{st} estimated, which indicates a relative tendency to zero with increase in sample size and number of measurements per subject, as expected for independent data. For very small sample sizes, however, PL's performance is compromised as can also be seen from Table 4.2 and Figure 4.2. Though θ_{st} hails from a Poisson distribution and is expected to be strictly positive, we argue that this interpretation takes effect in a hierarchical modeling framework. In the context of marginal modeling, this parameter can also take on negative values as is seen in Table 4.1. This phenomenon is often a source of confusion, and it is less well understood in non-Gaussian cases than for continuously distributed hierarchical data. Pryseley *et al.* (2011) describe how such negative correlations can be estimated and interpreted for both Gaussian and non-Gaussian settings. One important situation where negative association is natural is where cluster members are in a competitive relation with one another. Molenberghs and Verbeke (2011) further discuss how a negative correlation can be reconciled with a hierarhical model interpretation.

Simulations with θ_{st} strictly positive in the case of data with association, see Table A.1 in the appendix, in a marginal model perspective slightly improved the convergence rate while the bias and the MSE were more or less the same. In the case of data without association, the bias and MSE were also similar whether or not θ_{st} was constrained to be positive but the rate of convergence was reduced in the case of strictly positive θ_{st}.

In the presence of correlation, θ_{st} is estimated above 1 as can be seen from Table 4.3 for the various combinations of K and n_i. Convergence (Table 4.4) issues still persist for $K \approx 10$.

Data on epileptic seizures introduced in Section 2.2 were analyzed using this PL approach. A comparison with two other approaches, namely, (a) the standard Poisson regression assuming independence, and, (b) GEE1 were also used and results compared. Table 4.5 shows results of fitting a model for the evolution of the two treatment arms over time and, the same model but

Table 4.1: *Simulation study, no association: Parameter estimates for GEE (exch. correlation) and pseudo-likelihood for varying number of measurements per subject (n_i) and sample size (K)*

		GEE				Pseudo likelihood				
K	n_i	ξ_0	ξ_1	ξ_2	ξ_3	ξ_0	ξ_1	ξ_2	ξ_3	θ_{st}
True value		1.4531	-0.1869	-0.0328	0.0195	1.4531	-0.1869	-0.0328	0.0195	.
10	2	1.4147	-0.1691	-0.0306	0.0156	0.7424	-0.3586	-0.1007	0.1927	0.8545
10	4	1.4424	-0.1842	-0.0353	0.0203	1.4650	-0.1754	-0.0333	0.0178	-0.1502
10	8	1.4452	-0.1886	-0.0329	0.0203	1.4653	-0.1851	-0.0323	0.0199	-0.0906
10	16	1.4474	-0.1863	-0.0327	0.0196	1.4578	-0.1836	-0.0323	0.0193	-0.0465
100	2	1.4512	-0.1777	-0.0336	0.0122	1.4488	-0.1803	-0.0339	0.0122	-0.0066
100	4	1.4527	-0.1882	-0.0337	0.0201	1.4561	-0.1875	-0.0335	0.0200	-0.0179
100	8	1.4525	-0.1868	-0.0331	0.0197	1.4547	-0.1864	-0.0330	0.0197	-0.0101
100	16	1.4525	-0.1868	-0.0328	0.0197	1.4536	-0.1866	-0.0328	0.0197	-0.0051
1,000	2	1.4546	-0.1874	-0.0342	0.0200	1.4546	-0.1874	-0.0342	0.0199	-0.0016
1,000	4	1.4538	-0.1863	-0.0332	0.0193	1.4541	-0.1863	-0.0332	0.0193	-0.0015
1,000	8	1.4527	-0.1857	-0.0328	0.0194	1.4532	-0.1856	-0.0328	0.0194	-0.0021
1,000	16	1.4527	-0.1858	-0.0328	0.0194	1.4529	-0.1858	-0.0328	0.0194	-0.0005
10,000	2	1.4536	-0.1883	-0.0330	0.0202	1.4536	-0.1883	-0.0330	0.0202	0.0000
10,000	4	1.4536	-0.1865	-0.0329	0.0194	1.4537	-0.1865	-0.0329	0.0194	-0.0006
10,000	8	1.4536	-0.1875	-0.0329	0.0196	1.4536	-0.1875	-0.0329	0.0196	-0.0001
10,000	16	1.4531	-0.1871	-0.0328	0.0195	1.4531	-0.1871	-0.0328	0.0195	-0.0000

Table 4.2: *Simulation study, no association: Absolute bias in the parameter estimates and percent rate of convergence ($RATE_c$) for GEE and pseudo-likelihood for varying number of measurements per subject (n_i) and sample size (K)*

		GEE					Pseudo likelihood				
K	n_i	ξ_0	ξ_1	ξ_2	ξ_3	$RATE_c$	ξ_0	ξ_1	ξ_2	ξ_3	$RATE_c$
10	2	0.0384	0.0178	0.0022	0.0039	99	0.7107	0.1717	0.0679	0.1732	68
10	4	0.0107	0.0027	0.0025	0.0008	100	0.0119	0.0115	0.0005	0.0017	95
10	8	0.0079	0.0017	0.0001	0.0008	100	0.0122	0.0018	0.0005	0.0004	100
10	16	0.0057	0.0006	0.0001	0.0001	100	0.0047	0.0033	0.0005	0.0002	100
100	2	0.0019	0.0092	0.0008	0.0073	100	0.0043	0.0066	0.0011	0.0073	100
100	4	0.0004	0.0013	0.0009	0.0006	100	0.0030	0.0006	0.0007	0.0005	100
100	8	0.0006	0.0001	0.0003	0.0002	100	0.0016	0.0005	0.0002	0.0002	100
100	16	0.0006	0.0001	0.0000	0.0002	100	0.0005	0.0003	0.0000	0.0002	100
1,000	2	0.0015	0.0005	0.0014	0.0005	100	0.0015	0.0005	0.0014	0.0004	100
1,000	4	0.0007	0.0006	0.0004	0.0002	100	0.0010	0.0006	0.0004	0.0002	100
1,000	8	0.0004	0.0012	0.0000	0.0001	100	0.0001	0.0013	0.0000	0.0001	100
1,000	16	0.0004	0.0011	0.0000	0.0001	100	0.0002	0.0011	0.0000	0.0001	100
10,000	2	0.0005	0.0014	0.0002	0.0007	100	0.0005	0.0014	0.0002	0.0007	100
10,000	4	0.0005	0.0004	0.0001	0.0001	100	0.0006	0.0004	0.0001	0.0001	100
10,000	8	0.0005	0.0006	0.0001	0.0001	100	0.0005	0.0006	0.0001	0.0001	100
10,000	16	0.0000	0.0002	0.0000	0.0000	100	0.0000	0.0002	0.0000	0.0000	100

Table 4.3: *Simulation study, association: Parameter estimates of GEE (exch. correlation) and pseudo-likelihood for varying number of measurements per subject (n_i) and sample size (K)*

		GEE				Pseudo likelihood				
K	n_i	ξ_0	ξ_1	ξ_2	ξ_3	ξ_0	ξ_1	ξ_2	ξ_3	θ_{st}
True value		1.5807	-0.1881	-0.0340	0.0192	1.5807	-0.1881	-0.0340	0.0192	.
10	2	1.5183	-0.1862	-0.0470	0.0362	0.8550	-0.4958	-0.2284	0.3179	1.9756
10	4	1.5219	-0.1749	-0.0318	0.0222	1.2986	-0.2531	-0.0529	0.0343	1.5239
10	8	1.5328	-0.1765	-0.0329	0.0199	1.3454	-0.2226	-0.0473	0.0240	1.3882
10	16	1.5442	-0.1716	-0.0333	0.0199	1.4176	-0.2365	-0.0473	0.0256	1.2150
100	2	1.5651	-0.1834	-0.0311	0.0173	1.2493	-0.2745	-0.0561	0.0318	1.9167
100	4	1.5670	-0.1806	-0.0309	0.0191	1.2811	-0.2670	-0.0528	0.0282	1.8183
100	8	1.5703	-0.1882	-0.0326	0.0193	1.3179	-0.2612	-0.0520	0.0283	1.7083
100	16	1.5725	-0.1798	-0.0328	0.0195	1.3931	-0.2718	-0.0509	0.0284	1.4705
1,000	2	1.5774	-0.1853	-0.0328	0.0187	1.2591	-0.2709	-0.0535	0.0293	1.9175
1,000	4	1.5788	-0.1870	-0.0329	0.0192	1.2802	-0.2687	-0.0532	0.0283	1.8566
1,000	8	1.5776	-0.1867	-0.0329	0.0196	1.3179	-0.2713	-0.0524	0.0290	1.7337
1,000	16	1.5783	-0.1865	-0.0327	0.0195	1.3909	-0.2738	-0.0512	0.0287	1.4998
10,000	2	1.5779	-0.1872	-0.0326	0.0196	1.2628	-0.2712	-0.0549	0.0298	1.9194
10,000	4	1.5787	-0.1880	-0.0329	0.0198	1.2810	-0.2706	-0.0533	0.0289	1.8568
10,000	8	1.5778	-0.1863	-0.0328	0.0195	1.3179	-0.2715	-0.0525	0.0290	1.7367
10,000	16	1.5780	-0.1871	-0.0328	0.0195	1.3897	-0.2742	-0.0512	0.0287	1.5018

Table 4.4: *Simulation study, association: Absolute bias in the parameter estimates and percent rate of convergence ($RATE_c$) for GEE and pseudo-likelihood for varying number of measurements per subject (n_i) and sample size (K)*

		GEE					Pseudo likelihood				
K	n_i	ξ_0	ξ_1	ξ_2	ξ_3	$RATE_c$	ξ_0	ξ_1	ξ_2	ξ_3	$RATE_c$
10	2	0.0624	0.0019	0.0130	0.0170	85	0.7257	0.3077	0.1944	0.2987	97
10	4	0.0588	0.0132	0.0022	0.0030	100	0.2821	0.0650	0.0189	0.0151	100
10	8	0.0479	0.0116	0.0011	0.0007	100	0.2353	0.0345	0.0133	0.0048	100
10	16	0.0365	0.0165	0.0007	0.0007	100	0.1631	0.0484	0.0133	0.0064	100
100	2	0.0156	0.0047	0.0029	0.0019	100	0.3314	0.0864	0.0221	0.0126	100
100	4	0.0137	0.0075	0.0031	0.0001	100	0.2996	0.0789	0.0188	0.0090	100
100	8	0.0104	0.0001	0.0014	0.0001	100	0.2628	0.0731	0.0180	0.0091	99
100	16	0.0082	0.0083	0.0012	0.0003	100	0.1876	0.0837	0.0169	0.0092	99
1,000	2	0.0033	0.0028	0.0012	0.0005	100	0.3216	0.0828	0.0195	0.0101	100
1,000	4	0.0019	0.0011	0.0011	0.0000	100	0.3005	0.0806	0.0192	0.0091	99
1,000	8	0.0031	0.0014	0.0011	0.0004	100	0.2628	0.0832	0.0184	0.0098	99
1,000	16	0.0024	0.0016	0.0013	0.0003	100	0.1898	0.0857	0.0172	0.0095	97
10,000	2	0.0028	0.0009	0.0014	0.0004	100	0.3179	0.0831	0.0209	0.0106	98
10,000	4	0.0020	0.0001	0.0011	0.0006	100	0.2997	0.0825	0.0193	0.0097	97
10,000	8	0.0029	0.0018	0.0012	0.0003	100	0.2628	0.0834	0.0185	0.0098	98
10,000	16	0.0027	0.0010	0.0012	0.0003	100	0.1910	0.0861	0.0172	0.0095	98

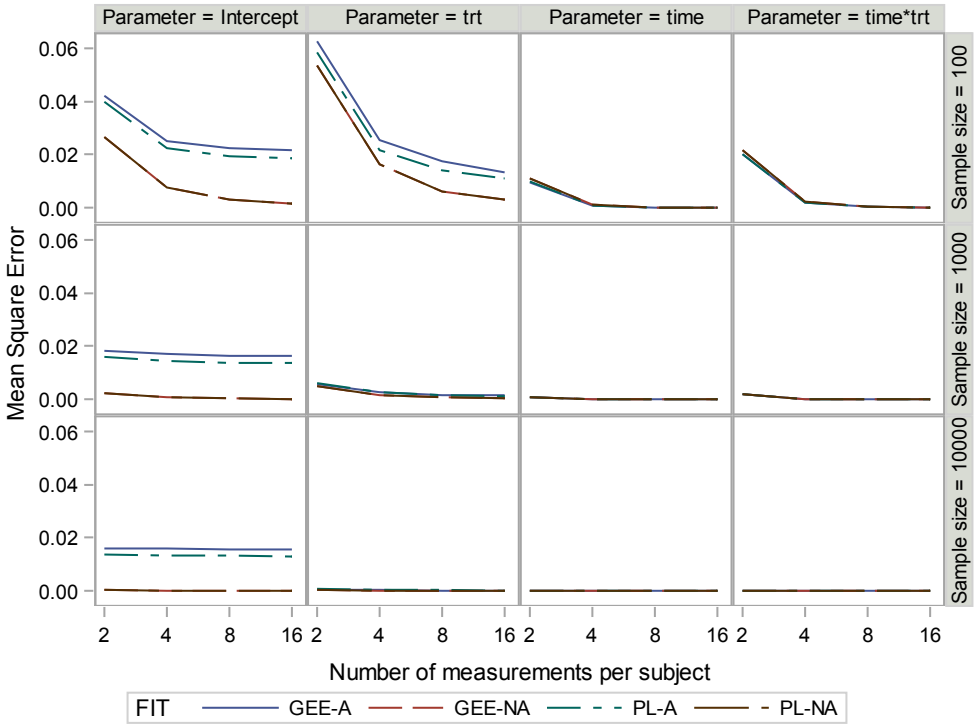


Figure 4.1: *Simulation study: Evolution of MSE by FIT over the number of measurements per subject, sample size $K = 10$ excluded. GEE-A refers to using GEE to model data with association while GEE-NA refers to using GEE to model data without association. Similarly, PL-A and PL-NA refer to using pseudo-likelihood to model data with or without association.*

with a correction for baseline characteristics of the patients. Similar results are observed for GEE and PL, especially as far as the standard errors are concerned.

4.4 Concluding Remarks

We have put forward a particular form of pseudo-likelihood, also termed pairwise likelihood, to estimate parameters for a model fitted to repeated count data. Beneficially, the specification of a bivariate count-data model only is required. Unlike conventional generalized estimating equations, our method

Table 4.5: *Epilepsy data: Parameter estimates (standard errors) for a univariate Poisson model, GEE (exchangeable correlation) and pseudo-likelihood (3.27). The first block refers to a model testing for a difference in number of epileptic seizures between the two treatment arms over time. The second block corrects for patient characteristics including race, age, sex, height and weight.*

Parameter	Univariate	GEE	Pseudo-likelihood
Intercept	1.4531 (0.0383)	1.3165 (0.1799)	0.91439 (0.29449)
treatment (0)	-0.1869 (0.0571)	0.0156 (0.2931)	-0.06423 (0.41424)
study week	-0.0328 (0.0038)	-0.0147 (0.0168)	-0.03891 (0.01875)
study week×treatment (0)	0.0195 (0.0058)	0.0035 (0.0201)	0.02845 (0.03558)
θ_{st}			1.10170 (0.26994)
Intercept	2.3963 (0.3576)	4.0954 (3.9610)	3.91804 (5.06465)
treatment (0)	-0.0992 (0.0578)	-0.0925 (0.2619)	-0.08047 (0.40335)
study week	-0.0299 (0.0039)	-0.0146 (0.0168)	-0.03403 (0.01824)
study week×treatment (0)	0.0168 (0.0058)	0.0033 (0.0206)	0.02247 (0.03298)
race (1)	-0.0811 (0.0506)	-0.3298 (0.2904)	-0.07743 (0.54786)
age (years)	-0.0188 (0.0017)	-0.0200 (0.0115)	-0.02025 (0.01936)
sex (1)	0.5747 (0.0575)	0.8959 (0.3936)	0.77549 (0.44018)
height	-0.0133 (0.0055)	-0.0429 (0.0576)	-0.03617 (0.07108)
weight	0.0008 (0.0005)	0.0023 (0.0040)	-0.00235 (0.00830)
θ_{st}			1.07935 (0.25153)

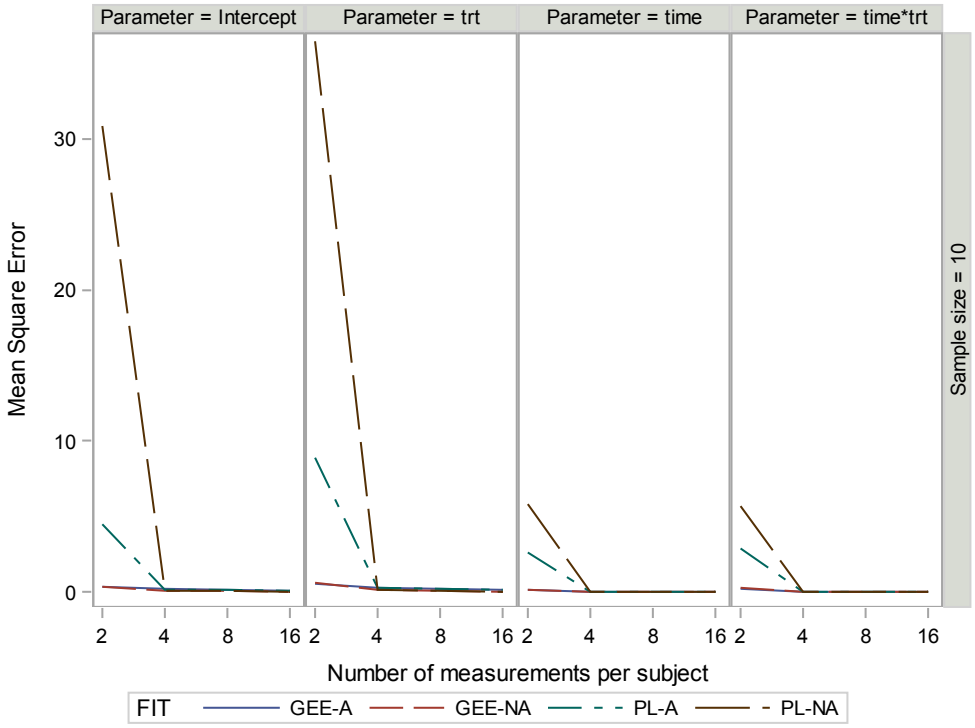


Figure 4.2: *Simulation study: Evolution of MSE by FIT over the number of measurements per subject for* $K = 10$. *GEE-A refers to using GEE to model data with association while GEE-NA refers to using GEE to model data without association. Similarly, PL-A and PL-NA refer to using pseudo-likelihood to model data with or without association.*

allows for the assessment of the association between pairs of measurements, in addition to the usual marginal mean parameters. Of course, one could consider a very general correlation structure with GEE1, but this cannot be subjected to standard statistical assessment, e.g., based on hypothesis-testing based assessment. Alternatively, one could switch to second-order GEE (Zhao and Prentice, 1990), but this may come with considerable computational complexity.

Pseudo-likelihood, like generalized estimating equations, yields consistent and asymptotically normally distributed parameter estimates with a sandwich estimator used to calculate the variance. On the one hand, GEE1 remains

computationally faster than PL because it only evaluates the first moment and plugs in working assumptions for the second. But because it allows for the miss-specification of the working correlation structure, one cannot rely on the correlation estimates from GEE1 for formulating answers to scientific questions, should interest be in the association as well. The computational burden encountered in PL grows with the number of measurements per subject or cluster size, as evaluation of the marginal PL is done for all $[n_i(n_i - 1)]/2$ possible pairs of a subject.

It is important to realize that the method used for simulation does not match the assumed model. This can be seen as a drawback, but underscores that more and more flexible methods for simulating correlated Poisson data are needed. It is a topic of ongoing research and Chapter 6 presents an alternative method for generating correlated counts.

The constant covariance terms, considered in this chapter, can and will be relaxed in future developments.

In conclusion, pseudo-likelihood is a viable alternative when pairwise association between repeated counts is of interest. Of course, while these pairwise association parameters are fully part of the model, in spite of the fact that full likelihood is not specified, there may be a price in terms of efficiency loss. At the same time, with pairwise pseudo-likelihood, no three-way of higher-order parameters can be estimated.

Further, and importantly, GEE2 and pairwise likelihood are less robust to misspecification of the association structure than conventional GEE. Of course, we have to place this against the background of functional restrictions on the correlation structure in marginal models. There are situations, especially with binary data, where a pairwise correlation structure is incompatible with the specified univariate mean functions. In such a case, it is better to have non-converging GEE1 and PL, than a converged GEE1 which nevertheless cannot correspond to a valid joint distribution.

Generally, the less parametric the model, the higher the robustness towards misspecification. This simply means that whatever is not specified, cannot be misspecified. In this spirit, PL is robust against the entire higher-order

association structure, given that it is not specified.

Robustness should also be seen against the existence of so-called parent distributions, i.e., full joint distributions that are compatible with the moments specified, e.g., the first and second moments in pairwise likelihood. Work has been done in this respect, e.g., by Molenberghs and Kenward (2010). These authors show that the parent provides a natural description of the framework into which the semi-parametrically specified parameters fit. The implication is that such semi-parametric methods as GEE1, GEE2, ALR, etc. can always be applied because there is always a valid parent, and hence a probabilistic basis. The sole condition is that the parametrically specified portion of the model be valid, but this is no different to any other statistical modeling exercise. It follows from the above that, when the pairwise correlation structure is grossly misspecified, the pairwise probabilities may be jeopardized and more so the parent distribution. This implies that robustness can come with important drawbacks. In pairwise likelihood, the modeler's obligation to reflect carefully on all that is specified is straightforwardly built in.

Chapter 5

Second-order Generalized Estimating Equations for Correlated Count Data

5.1 Introduction

Count data, as the name suggests, arises as a result of a counting process in a given interval of time and therefore takes on non-negative integer values. Examples may include: number of doctor visits, number of epileptic seizures, number of accidents, etc. To draw inferences from such data, a Poisson distribution is usually assumed as the data generating mechanism and a log-likelihood function is constructed which, when maximized, yields parameters of scientific interest. The standard Poisson model implies that the mean and variance are equal (McCullagh and Nelder, 1989). However, in practice, this implication is usually restrictive because count data samples often have the mean either greater than the variance (so-called underdispersion) or less than the variance (also known as overdispersion). Therefore, using the Poisson model in its basic form would not account for this feature correctly. To account for overdispersion, the negative-binomial (NB, Breslow, 1984; Lawless, 1987) model is an option. Also, count data regularly has an incidence of zero counts greater than expected from the Poisson model. The zero-inflated Poisson (ZIP, Lambert, 1992) or zero-inflated negative binomial (ZINB, Ridout

et al., 2001) model account for the extra zeros.

Further, and as mentioned in Section 3.3, count data is often collected repeatedly over time in many studies. Such studies aim at describing, for example, the evolution of the subjects' condition over time, given certain characteristics of interest. This repetition in the observation of the patients or cluster or subjects induces the aspect of correlation because responses from the same subject will be more alike than those between different subjects. Also here, extensions from cross-sectional or univariate data to correlated data have been proposed in the literature and implemented in statistical software packages. Some of these include generalized estimating equations (GEE1, Liang and Zeger, 1986), the Poisson-normal model, which belongs to the generalized linear mixed model family (GLMM, Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) or more generally the combined model (Molenberghs *et al.*, 2007, 2010), the multivariate negative binomial model (Solis-Trapala and Farewell, 2005; Winkelmann, 2008), etc. Research in modeling hierarchical or correlated count data is certainly ongoing. Iddi and Molenberghs (2013) also contributed to this area of correlated and overdispersed count data by proposing a marginalized model for zero-inflated, overdispersed, and correlated count data. We refer interested readers in the topic of generalized estimating equations to Molenberghs and Verbeke (2005), Hardin and Hilbe (2003), Diggle *et al.* (2002), Fitzmaurice *et al.* (2004), and Ziegler (2011).

In this chapter, marginal models are of interest, especially motivated by two datasets presented in Sections 2.3 and 2.4, and analyzed in Section 5.2. We hence-forth limit our discussion to the marginal-models framework for correlated count data. GEE, described in Section 3.3.6 are a common tool used when modeling correlated count data. A somewhat different route is taken here in that estimating equations are proposed at the level of subject i's pair of responses (Y_{is}, Y_{it}). The use of pairs rather than the whole vector of responses $(\boldsymbol{Y}_i)$ would lead to loss of efficiency in estimating the parameters of interest but would simplify the computational unattractiveness of having to obtain the third and fourth moments that is evident as long as scientific interest lies in both the marginal mean parameters $(\boldsymbol{\xi})$ and the association

structure. Section 5.1.1 presents our extension of the estimating equations to model the covariance structure via covariates simultaneously with the marginal mean parameters by incorporating the bivariate Poisson distribution into the estimating equations permitting inference on both the marginal mean and covariance parameters.

5.1.1 Extension of GEE using the Bivariate Poisson Distribution

To put matters into perspective, consider the following bivariate Poisson distribution which is derived using the trivariate reduction method (Kocherlakota and Kocherlakota, 1992, 2001) based on a convolution of independent Poisson variables. Note that there are several derivations of the bivariate Poisson distribution in the literature. For example, Lakshminarayana *et al.* (1999) derive their bivariate Poisson distribution based on a polynomial factor. Assume that W_{ic} are independent Poisson random variables such that $\mathsf{E}(W_{ic}) = \eta_{ic}$, $c = s, t$ or st. The random variables $Y_{is} = (W_{is} + W_{ist})$ and $Y_{it} = (W_{it} + W_{ist})$ then follow a bivariate Poisson distribution. Thus, $(Y_{is}, Y_{it}) \sim BP(\eta_{is}, \eta_{it}, \eta_{ist})$ characterized by

$$f(y_{is}, y_{it}) = e^{-(\eta_{is}+\eta_{it}+\eta_{ist})} \frac{\eta_{is}^{y_{is}}}{y_{is}!} \frac{\eta_{it}^{y_{it}}}{y_{it}!} \sum_{l=0}^{\min(y_{is},y_{it})} \binom{y_{is}}{l} \binom{y_{it}}{l} l! \left(\frac{\eta_{ist}}{\eta_{is}\eta_{it}}\right)^{l}. \quad (5.1)$$

Marginally, $\mathsf{E}(Y_{is}) = \eta_{is} + \eta_{ist}$, $\mathsf{E}(Y_{it}) = \eta_{it} + \eta_{ist}$ and $\mathsf{Cov}(Y_{is}, Y_{it}) = \eta_{ist}$. We propose the score equation $(U_{i,st})$ to be computed at each pair $\{s, t\}$ of responses from subject i such that the estimates for the $\boldsymbol{\xi}$ regression parameters are obtained by solving

$$\sum_{i}^{K} U_i = \sum_{i}^{K} \sum_{1 \le s < t \le n_i} U_{i,st}(\boldsymbol{\xi}) = \sum_{i}^{K} \sum_{s<t} \frac{\partial \boldsymbol{\mu}_{i,st}}{\partial \boldsymbol{\xi}^{\top}} V_{i,st}^{-1} \left(\boldsymbol{Y}_{i,st} - \boldsymbol{\mu}_{i,st}\right) = 0, \quad (5.2)$$

where

$$\boldsymbol{Y}_{i,st} = \begin{pmatrix} Y_{is} \\ Y_{it} \\ Y_{is}Y_{it} \end{pmatrix}, \ \boldsymbol{\mu}_{i,st} = \begin{pmatrix} \mathsf{E}(Y_{is}) \\ \mathsf{E}(Y_{it}) \\ \mathsf{E}(Y_{is}Y_{it}) \end{pmatrix} \text{ and}$$

$$V_{i,st} = \begin{pmatrix} \mathsf{Var}(Y_{is}) & \mathsf{Cov}(Y_{is}, Y_{it}) & \mathsf{Cov}(Y_{is}, Y_{is}Y_{it}) \\ \mathsf{Cov}(Y_{it}, Y_{is}) & \mathsf{Var}(Y_{it}) & \mathsf{Cov}(Y_{it}, Y_{is}Y_{it}) \\ \mathsf{Cov}(Y_{is}Y_{it}, Y_{is}) & \mathsf{Cov}(Y_{is}Y_{it}, Y_{it}) & \mathsf{Var}(Y_{is}Y_{it}) \end{pmatrix},$$

with $\mathsf{E}(Y_{is}Y_{it}) = \mathsf{E}(Y_{is})\mathsf{E}(Y_{it}) + \mathsf{Cov}(Y_{is}, Y_{it})$, $\mathsf{Var}(Y_{is}) = \mathsf{E}(Y_{is})$, $\mathsf{Var}(Y_{it}) = \mathsf{E}(Y_{it})$ and $\mathsf{E}(Y_{is}Y_{it}) = \mathsf{E}(Y_{is})\mathsf{E}(Y_{it}) + \eta_{ist}$.

To derive the covariance terms $\mathsf{Cov}(Y_{is}, Y_{is}Y_{it})$, $\mathsf{Cov}(Y_{it}, Y_{is}Y_{it})$ and $\mathsf{Var}(Y_{is}Y_{it})$ in V_{ist} in (5.2), we need to calculate the following four moments of the Poisson distribution that turn out to be essential. If $\tilde{X} \sim \text{Poisson}(\lambda)$ with probability mass function $f(\tilde{X}; \lambda) = \frac{\lambda^{\tilde{x}} e^{-\lambda}}{\tilde{x}!}$ where $\lambda > 0$ and $\tilde{x} = 0, 1, 2, \ldots$, then the n^{th} moment $\mathsf{E}(\tilde{X}^n)$, $n = 1, 2, 3, 4$ is as follows;

$$\begin{aligned}
\mathsf{E}(\tilde{X}) &= e^{-\lambda} \sum_{\tilde{x}=0}^{\infty} \tilde{x} \frac{\lambda^{\tilde{x}}}{x!} = \lambda e^{-\lambda} \sum_{\tilde{x}=1}^{\infty} \frac{\lambda^{\tilde{x}-1}}{(\tilde{x}-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda, \\
\mathsf{E}(\tilde{X}^2) &= e^{-\lambda} \sum_{\tilde{x}=0}^{\infty} \tilde{x}^2 \frac{\lambda^{\tilde{x}}}{\tilde{x}!} = \lambda e^{-\lambda} \sum_{\tilde{x}=1}^{\infty} \tilde{x} \frac{\lambda^{\tilde{x}-1}}{(\tilde{x}-1)!} \\
&= \lambda e^{-\lambda} \frac{\partial}{\partial \lambda} \left[\lambda \sum_{\tilde{x}=1}^{\infty} \frac{\lambda^{\tilde{x}-1}}{(\tilde{x}-1)!} \right] = \lambda e^{-\lambda} \frac{\partial}{\partial \lambda} \left(\lambda e^{\lambda} \right) \\
&= \lambda e^{-\lambda} (1 + \lambda) e^{\lambda} = \lambda (1 + \lambda), \\
\mathsf{E}(\tilde{X}^3) &= e^{-\lambda} \sum_{\tilde{x}=0}^{\infty} \tilde{x}^3 \frac{\lambda^{\tilde{x}}}{\tilde{x}!} = \lambda e^{-\lambda} \sum_{\tilde{x}=1}^{\infty} \tilde{x}^2 \frac{\lambda^{\tilde{x}-1}}{(\tilde{x}-1)!} \\
&= \lambda e^{-\lambda} \frac{\partial}{\partial \lambda} \left[\lambda \sum_{\tilde{x}=1}^{\infty} \frac{\tilde{x} \lambda^{\tilde{x}-1}}{(\tilde{x}-1)!} \right] \\
&= \lambda e^{-\lambda} \frac{\partial}{\partial \lambda} \left[\lambda \frac{\partial}{\partial \lambda} \left(\lambda \sum_{\tilde{x}=1}^{\infty} \frac{\lambda^{\tilde{x}-1}}{(\tilde{x}-1)!} \right) \right] \\
&= \lambda e^{-\lambda} \frac{\partial}{\partial \lambda} \left[\lambda \frac{\partial}{\partial \lambda} \left(\lambda e^{\lambda} \right) \right] = \lambda e^{-\lambda} \frac{\partial}{\partial \lambda} \left(\lambda (1 + \lambda) e^{\lambda} \right) \\
&= \lambda e^{-\lambda} (\lambda (1 + \lambda) + 1 + 2\lambda) e^{\lambda}) = \lambda (1 + 3\lambda + \lambda^2)
\end{aligned} \tag{5.3}$$

and

$$\begin{aligned}\mathsf{E}(\tilde{X}^4) &= \lambda e^{-\lambda}\frac{\partial}{\partial\lambda}\left[\lambda(1+3\lambda+\lambda^2)e^{\lambda}\right] \\ &= \lambda\left[\lambda(1+3\lambda+\lambda^2)+1+6\lambda+3\lambda^2\right] = \lambda(\lambda^3+6\lambda^2+7\lambda+1).\end{aligned} \tag{5.4}$$

Generally, if $\mathsf{E}(\tilde{X}^n) = f_n(\lambda)$, then

$$\begin{aligned}f_{n+1}(\lambda) &= \lambda e^{-\lambda}\frac{\partial}{\partial\lambda}\left[f_n(\lambda)e^{\lambda}\right] \\ &= \lambda e^{-\lambda}\left[f_n^{'}(\lambda)+f_n(\lambda)\right]e^{\lambda} = \lambda\left[f_n^{'}(\lambda)+f_n(\lambda)\right],\end{aligned} \tag{5.5}$$

where $f_n^{'}(\cdot)$ is the first derivative of $f_n(\cdot)$. Now, from $\mathsf{Var}\left(\tilde{X}\right) = \mathsf{E}\left(\tilde{X}^2\right) - \left[\mathsf{E}\left(\tilde{X}\right)\right]^2$, it follows that $\mathsf{Var}\left(Y_{is}Y_{it}\right) = \mathsf{E}\left[(Y_{is}Y_{it})^2\right] - \left[\mathsf{E}(Y_{is}Y_{it})\right]^2$ such that $\mathsf{E}\left[(Y_{is}Y_{it})^2\right]$ or $\mathsf{E}(Y_{is}^2Y_{it}^2)$ is to be replaced with

$$\begin{aligned}\mathsf{E}(Y_{is}^2Y_{it}^2) &= \mathsf{E}\left[\left(W_{is}+W_{ist}\right)^2\left(W_{it}+W_{ist}\right)^2\right] \\ &= \mathsf{E}\left[\left(W_{is}^2+2W_{is}W_{ist}+W_{ist}^2\right)\left(W_{it}^2+2W_{it}W_{ist}+W_{ist}^2\right)\right] \\ &= \mathsf{E}\left[\begin{array}{l} W_{is}^2W_{it}^2+2W_{is}^2W_{it}W_{ist}+W_{is}^2W_{ist}^2+ \\ 2W_{is}W_{it}^2W_{ist}+4W_{is}W_{it}W_{ist}^2+2W_{is}W_{ist}^3+ \\ W_{it}^2W_{ist}^2+2W_{it}W_{ist}^3+W_{ist}^4 \end{array}\right],\end{aligned} \tag{5.6}$$

where further simplification is possible by applying the expectation to the independent Poisson variables W_s, W_t, W_{st} and using the moments in (5.3)-(5.4). This leads to the solution

$$\begin{aligned}\mathsf{E}(Y_{is}^2Y_{it}^2) = \mathsf{E}(Y_{is})^2\mathsf{E}(Y_{it})^2 + \mathsf{E}(Y_{it})\mathsf{E}(Y_{is})^2) + \mathsf{E}(Y_{is})\mathsf{E}(Y_{it})^2 + 2\eta_{ist}^2 + \\ \mathsf{E}(Y_{is})\mathsf{E}(Y_{it})(1+4\eta_{ist}) + 2\eta_{ist}(\mathsf{E}(Y_{is})+\mathsf{E}(Y_{it})) + \eta_{ist}.\end{aligned} \tag{5.7}$$

The covariances $\mathsf{Cov}(Y_{is}, Y_{is}Y_{it})$ and $\mathsf{Cov}(Y_{is}, Y_{is}Y_{it})$ are calculated as

$$\mathsf{Cov}(Y_{is}, Y_{is}Y_{it}) = \mathsf{E}(Y_{is}Y_{is}Y_{it}) - \mathsf{E}(Y_{is})\mathsf{E}(Y_{is}Y_{it}) = \mathsf{E}(Y_{is}^2Y_{it}) - \mathsf{E}(Y_{is})\mathsf{E}(Y_{is}Y_{it})$$

and

$$\mathsf{Cov}(Y_{it}, Y_{is}Y_{it}) = \mathsf{E}(Y_{it}Y_{is}Y_{it}) - \mathsf{E}(Y_{it})\mathsf{E}(Y_{is}Y_{it}) = \mathsf{E}(Y_{is}Y_{it}^2) - \mathsf{E}(Y_{it})\mathsf{E}(Y_{is}Y_{it}),$$

respectively, where similar algebra as in (5.6) leads to the following quantities:

$$\begin{aligned}
\mathsf{E}(Y_{is}^2) &= \mathsf{E}(Y_{is}) + \mathsf{E}(Y_{is})^2, \\
\mathsf{E}(Y_{it}^2) &= \mathsf{E}(Y_{it}) + \mathsf{E}(Y_{it})^2, \\
\mathsf{E}(Y_{is}^2 Y_{it}) &= \mathsf{E}(Y_{is}^2)\mathsf{E}(Y_{it}) + 2\eta_{ist}\mathsf{E}(Y_{is}) + \eta_{ist}, \\
\mathsf{E}(Y_{is}Y_{it}^2) &= \mathsf{E}(Y_{is})\mathsf{E}(Y_{it}^2) + 2\eta_{ist}\mathsf{E}(Y_{it}) + \eta_{ist}.
\end{aligned}$$

The means $\mathsf{E}(W_{ic})$, are related to covariates as $\log\left[\mathsf{E}(W_{ic})\right] = \mathbf{X}_{ic}^{\top}\boldsymbol{\xi}$, where

$$\begin{pmatrix} \boldsymbol{X}_{is}^{\top} \\ \boldsymbol{X}_{it}^{\top} \\ \boldsymbol{X}_{ist}^{\top} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & X_{is1} & X_{is2} & \dots & X_{isp} \\ 0 & 1 & 0 & X_{it1} & X_{it2} & \dots & X_{itp} \\ 0 & 0 & 1 & X_{ist1} & X_{ist2} & \dots & X_{istp} \end{pmatrix}. \tag{5.8}$$

The vector of unknown regression parameters $\boldsymbol{\xi} = (\xi_{0s}, \xi_{0t}, \xi_{0st}, \xi_1, \xi_2, \dots, \xi_p)^{\top}$, allowing for an intercept ξ_0 specific to time point s, t and their product st in addition to the regression parameters $(\xi_1, \dots, \xi_p)$ shared for the rest of the variables. Furthermore, the model-based standard errors are obtained as the square root of the diagonal entries of

$$U^* = \left(\sum_{i}^{K} \sum_{s<t} \frac{\partial \boldsymbol{\mu}_{i,st}}{\partial \boldsymbol{\xi}^{\top}} V_{i,st}^{-1} \frac{\partial \boldsymbol{\mu}_{i,st}}{\partial \boldsymbol{\xi}} \right)^{-1},$$

while the sandwich standard errors are calculated as the square root of the diagonal of

$$U^{**} = U^* \cdot I^* \cdot U^* = U^* \cdot \sum_{i}^{K} \sum_{s<t} U_{i,st} U_{i,st}^{\top} \cdot U^*.$$

As mentioned in Section 3.3.1, covariates under consideration in (5.8) may be either time-stationary or time-varying. When time-stationary, each of the

columns $1, 2, \ldots, p$ would contain the same values. On the other hand, when time-varying, X_{ist} can be derived as a function of X_{is} and X_{it}, for example, a difference, lag, sum, product, ratio, etc. The correlation between two measurements Y_{is} and Y_{it} is then calculated as $\rho_{i,st} = \mathsf{Cov}(Y_{is}, Y_{it})/\sqrt{\mathsf{Var}(Y_{is})\mathsf{Var}(Y_{it})}$.

5.2 Data Analysis

To analyze the epilepsy dataset presented in Section 2.4, the following covariates are considered: baseline (the 8-week pre-randomization seizure count), age (years), treatment and time (visit), denoted as $B^\star$, $A^\star$, $T^\star$ $t^\star$, respectively. As mentioned in Section 2.4, one patient is observed in Figure 2.6 to have a seemingly outlying profile. However, Thall and Vail (1990) find no clinical basis to tag the patient as an extreme case. The following analyses of the Epilepsy data therefore use data for all the 59 patients in the study.

Considering no time-varying covariates for the covariance $\mathsf{E}(W_{ist}) = \eta_{ist}$, we fitted the model

$$\begin{aligned}
\log(\eta_{is}) &= \xi_{0s} + \xi_1 t^\star_{is} + \xi_2 T^\star_i + \xi_3 (t^\star_{is} \times T^\star_i) + \xi_4 B^\star_i + \xi_5 A^\star_i, \\
\log(\eta_{it}) &= \xi_{0t} + \xi_1 t^\star_{it} + \xi_2 T^\star_i + \xi_3 (t^\star_{it} \times T^\star_i) + \xi_4 B^\star_i + \xi_5 A^\star_i, \\
\log(\eta_{ist}) &= \xi_{0st} + \xi_2 T^\star_i + \xi_5 A^\star_i.
\end{aligned} \tag{5.9}$$

In general, the covariates used for modeling $\log(\eta_{ist})$ may be the same or different from those used for η_{is} and η_{it}. Table 5.1 shows the parameter estimates and standard errors corresponding to Model 5.9. The model with time-varying covariate *visit* is

$$\begin{aligned}
\log(\eta_{is}) &= \xi_{0s} + \xi_1 t^\star_{is} + \xi_2 T^\star_i + \xi_3 (t^\star_{is} \times T^\star_i) + \xi_4 B^\star_i + \xi_5 A^\star_i, \\
\log(\eta_{it}) &= \xi_{0t} + \xi_1 t^\star_{it} + \xi_2 T^\star_i + \xi_3 (t^\star_{it} \times T^\star_i) + \xi_4 B^\star_i + \xi_5 A^\star_i, \\
\log(\eta_{ist}) &= \xi_{0st} + \xi_1 \varphi(t^\star_{is}, t^\star_{it}) + \xi_2 T^\star_i + \xi_5 A^\star_i,
\end{aligned} \tag{5.10}$$

where $\varphi(t^\star_{is}, t^\star_{it})$ denotes a function applied to the time-varying covariate *time*, in this case, a difference between *time* at point s and t $(t^\star_{is} - t^\star_{it})$. Other possibilities for $\varphi(\cdot)$ may include, for example, the lag, ratio, sum, product, etc.

From Table 5.1, the interaction between the visits and treatments is not significant ($p = 0.7977$) despite the fact that the mean profiles in Figure 2.8 suggest otherwise. The discrepancy between the observation in Section 2.4 and this finding is related to that one patient whose profile seems more extreme relative to the others. Mean profiles based on data without this potentially outlying patient (not shown) also suggested no interaction between the treatments and visits. Considering age as a time-varying covariate when modeling the covariance in (5.10) changed the results slightly but the conclusions remained similar to when only time stationary covariates are used to model the covariance. Table 5.4 shows the intervals of the minimum and maximum correlations, for the placebo and progabide groups, obtained from fitting (5.9) and (5.10). The correlations range over a wide interval with the progabide group having even wider ranges. By modeling the covariance between two measurements using visit as a time-varying covariate and a difference as the time-varying function seems to have a minor impact on the parameter estimates but also on correlations.

To analyze the Jimma study introduced in Section 2.3, denote the covariates age, sex, breastfeeding, help, rural place and semi-urban place as $a^\star$, $s^\star$, $b^\star$, $h^\star$, $p_1^\star$ and $p_2^\star$, respectively. The model fitted considering *age* as the time-varying covariate is

$$\begin{aligned}
\log\left[\mathsf{E}(W_{is})\right] &= \xi_{0s} + \xi_1 a_{is}^\star + \xi_2 s_i^\star + \xi_3 b_i^\star + \xi_4 h_i^\star + \xi_5 p_{1i}^\star + \xi_6 p_{2i}^\star, \\
\log\left[\mathsf{E}(W_{it})\right] &= \xi_{0t} + \xi_1 a_{it}^\star + \xi_2 s_i^\star + \xi_3 b_i^\star + \xi_4 h_i^\star + \xi_5 p_{1i}^\star + \xi_6 p_{2i}^\star, \\
\log\left[\mathsf{E}(W_{ist})\right] &= \xi_{0st} + \xi_1 \varphi(a_{is}^\star, a_{it}^\star) + \xi_2 s_i^\star + \xi_3 b_i^\star + \xi_4 h_i^\star,
\end{aligned} \tag{5.11}$$

where $\varphi(a_{is}^\star, a_{it}^\star) = a_{is}^\star - a_{it}^\star$. Table 5.3 confirms the observations made in Section 2.3, namely, that that the average number of days of diarrheal illness increases as the infants grow older with the female infants having lower counts of days compared to the males. We also find that not breastfeeding is positively related to the number of days of diarrheal illness ($p = 0.0150$) while not seeking medical help is also highly statistically signicant in increasing the number of days of having diarrhea in the infants. Table 5.5 shows the minimum and maximum estimates of the correlation by gender obtained from fitting (5.11).

The ranges of the correlation are a bit narrower than those from the epilepsy but there are minor differences in the correlation estimates between males and females. Unlike the Epilepsy data case where the minimum and maximum estimates of the correlation seem not to change much over time, the Jimma dataset reflects a decreasing trend in the correlations as the infants get older, in the sense that measurements close together are more correlated than those further apart.

5.3 Concluding Remarks

In this chapter, we have worked on estimating equations that can be used for modeling longitudinal data with the goal of making inference on (sub)populations. These estimating equations model the dependence of the mean response on covariates of interest, without specifying the joint distribution of the vector of responses from a subject. Should scientific interest lie only in the estimation of the so-called population averaged parameters, the approach of Liang and Zeger (1986) is quite sufficient and one need not worry about more involving methods. Because in practice, the method of Liang and Zeger (1986) is limited should interest lie also in the association structure, alternatives have been proposed. For example, Prentice (1988) proposed simultaneous estimation of the marginal mean and association structure permitting inference also on the parameters characterizing the association, in the context of binary data. As has been shown in this chapter, the binary case is special as the model for the association is fully determined by the mean and covariance. For count data, however, this issue is a bit more involved and proposed solutions such as in Prentice and Zhao (1991) come to the rescue. They estimate the parameters of the marginal mean and association simultaneously without making the orthogonality assumption made by Zhao and Prentice (1990). This, however, is computationally unbecoming since it involves third- and higher-order moments.

We have presented estimating equations at pair level of the vector of responses for each subject in the context of correlated count data. The proposal

incorporates the bivariate Poisson distribution which allows the modeling of the covariance between two measurements. It is formulated such that the variance-covariance matrix of the outcome variable is not a nuisance but one on which inference can be made while the standard errors are estimated using a sandwich estimator. The method allows for time-stationary as well as time-varying covariates and gives the user the flexibility to determine which function to use for the time-varying covariates. Possibilities may be a lag, ratio, difference, sum, product, etc. A SAS macro has been written to implement this method and is available at `http://ibiostat.be/software/longitudinal` or `http://ibiostat.be/software/count`. Using a 64-bit Windows 8.1 operating system computer with 8GB RAM and 2.80GHz processor, (5.9) converged, based on a dataset of 236 observations, after 14 iterations with a real time of 0.45 seconds. Similarly, (5.10) took 0.38 seconds (real time) and converged after 14 iterations. Finally, (5.11) was fitted on a dataset of about $46,000$ observations and converged after 13 iterations and 59.39 seconds.

Table 5.1: *Epilepsy data: Parameter estimates and standard errors when a time stationary covariate is considered (Model 5.9).*

		Model-based			Sandwich or empirical		
Effect	Est.	s.e.	χ^2	p	s.e.	χ^2	p
Intercept (ξ_{0s})	−1.1904	0.2157	30.45	<.0001	1.2655	0.88	0.3468
Intercept (ξ_{0t})	−1.2119	0.2474	24.00	<.0001	1.2563	0.93	0.3347
Intercept (ξ_{0st})	1.3148	0.0929	200.17	<.0001	0.4836	7.39	0.0066
visit (ξ_1)	−0.1724	0.0396	18.95	<.0001	0.0500	11.87	0.0006
trt(Placebo) (ξ_2)	0.2782	0.0452	37.89	<.0001	0.2490	1.25	0.2638
trt*visit (ξ_3)	0.0404	0.0390	1.07	0.3004	0.1576	0.07	0.7977
Baseline (ξ_4)	0.0379	0.0015	629.29	<.0001	0.0088	18.46	<.0001
Age (ξ_5)	0.0082	0.0027	8.93	0.0028	0.0129	0.40	0.5257

Table 5.2: *Epilepsy data: Parameter estimates and standard errors when a time-varying covariate is considered (Model 5.10).*

		Model-based			Sandwich or empirical		
Effect	Est.	s.e.	χ^2	p	s.e.	χ^2	p
Intercept (ξ_{0s})	−1.3907	0.2221	39.20	<.0001	1.3798	1.02	0.3135
Intercept (ξ_{0t})	−1.6024	0.2518	40.50	<.0001	1.4621	1.20	0.2731
Intercept (ξ_{0st})	1.2180	0.0992	150.64	<.0001	0.5099	5.71	0.0169
visit (ξ_1)	−0.0457	0.0246	3.45	0.0633	0.0468	0.95	0.3289
trt(Placebo) (ξ_2)	0.3037	0.0448	46.04	<.0001	0.2456	1.53	0.2163
trt*visit (ξ_3)	0.0066	0.0375	0.03	0.8603	0.1550	0.00	0.9660
Baseline (ξ_4)	0.0378	0.0015	619.46	<.0001	0.0091	17.27	<.0001
Age (ξ_5)	0.0085	0.0027	9.72	0.0018	0.0129	0.44	0.5094

Table 5.3: *Jimma data: Parameter estimates and standard errors when a time-varying covariate is considered (Model 5.11).*

		Model-based			Sandwich or empirical		
Effect	Est.	s.e.	χ^2	p	s.e.	χ^2	p
Intercept (ξ_{0s})	−1.3373	0.0090	21968.71	<.0001	0.2232	35.89	<.0001
Intercept (ξ_{0t})	−0.9529	0.0098	10068.03	<.0001	0.1791	28.31	<.0001
Intercept (ξ_{0st})	−0.3984	0.0091	3006.32	<.0001	0.1775	5.04	0.0248
age (ξ_1)	0.1260	0.0007	28725.74	<.0001	0.0083	228.11	<.0001
sex(Female) (ξ_2)	−0.1732	0.0034	2556.79	<.0001	0.0307	31.93	<.0001
bf(No) (ξ_3)	0.2379	0.0099	575.10	<.0001	0.0978	5.92	0.0150
help(No) (ξ_4)	2.0924	0.0038	308352.00	<.0001	0.0380	3034.68	<.0001
place(Rural) (ξ_5)	−0.0448	0.0064	48.82	<.0001	0.1060	0.18	0.6729
place(Semi-urban)(ξ_6)	−0.2088	0.0077	744.55	<.0001	0.1095	3.63	0.0566

Table 5.4: *Epilepsy data: Minimum and maximum correlations from fitting Model 5.9 (top panel) and Model 5.10 (bottom panel) for the two treatments.*

	Placebo				Progabide			
Visit	1	2	3	4	1	2	3	4
1	[1.00,1.00]				[1.00,1.00]			
2	[0.18,0.92]	[1.00,1.00]			[0.05,0.92]	[1.00,1.00]		
3	[0.19,0.93]	[0.20,0.93]	[1.00,1.00]		[0.05,0.93]	[0.06,0.94]	[1.00,1.00]	
4	[0.20,0.93]	[0.22,0.94]	[0.23,0.94]	[1.00,1.00]	[0.06,0.93]	[0.06,0.94]	[0.07,0.95]	[1.00,1.00]
1	[1.00,1.00]				[1.00,1.00]			
2	[0.20,0.93]	[1.00,1.00]			[0.05,0.93]	[1.00,1.00]		
3	[0.21,0.93]	[0.21,0.93]	[1.00,1.00]		[0.06,0.93]	[0.06,0.93]	[1.00,1.00]	
4	[0.22,0.94]	[0.22,0.94]	[0.21,0.93]	[1.00,1.00]	[0.06,0.94]	[0.06,0.94]	[0.06,0.93]	[1.00,1.00]

Table 5.5: *Jimma data: Minimum and maximum correlations from fitting Model 5.11 by gender.*

Gender	Age	N	0	2	4	6	8	10	12
Female	0	3453	[1.00,1.00]						
	2	3453	[0.27,0.81]	[1.00,1.00]					
	4	3336	[0.18,0.76]	[0.21,0.77]	[1.00,1.00]				
	6	3255	[0.14,0.69]	[0.16,0.71]	[0.18,0.72]	[1.00,1.00]			
	8	3174	[0.11,0.62]	[0.12,0.64]	[0.15,0.66]	[0.17,0.69]	[1.00,1.00]		
	10	3102	[0.08,0.53]	[0.09,0.56]	[0.10,0.58]	[0.12,0.60]	[0.13,0.62]	[1.00,1.00]	
	12	2667	[0.06,0.40]	[0.06,0.47]	[0.07,0.50]	[0.08,0.52]	[0.09,0.54]	[0.10,0.56]	[1.00,1.00]
Male	0	3549	[1.00,1.00]						
	2	3549	[0.27,0.81]	[1.00,1.00]					
	4	3411	[0.19,0.76]	[0.21,0.77]	[1.00,1.00]				
	6	3303	[0.14,0.69]	[0.16,0.71]	[0.20,0.72]	[1.00,1.00]			
	8	3219	[0.10,0.62]	[0.12,0.64]	[0.13,0.66]	[0.15,0.67]	[1.00,1.00]		
	10	3135	[0.08,0.53]	[0.09,0.56]	[0.10,0.60]	[0.11,0.60]	[0.13,0.62]	[1.00,1.00]	
	12	2721	[0.06,0.40]	[0.07,0.47]	[0.08,0.50]	[0.09,0.52]	[0.10,0.54]	[0.11,0.56]	[1.00,1.00]

Chapter 6

The Combined Model: A Tool for Simulating Correlated Counts with Overdispersion

6.1 Introduction

Research today generates a lot of data that have to be analyzed and summarized into meaningful and informative statements. Analysis is done using statistical methods that depend on the kind of data at hand. In medical research, it is often the case that data on a patient is profiled longitudinally in the sense that each patient is followed repeatedly or observed at multiple points over time. This introduces the phenomenon of correlated data because observations from one patient will be more related or similar than observations across different patients. A lot of research has already been committed to the analysis of correlated data. For example, Molenberghs and Verbeke (2005) and Verbeke and Molenberghs (2000) focus on methods for the analysis of discrete and continuous longitudinal data, respectively. In the context of continuous or normal longitudinal data, calculations are computationally easier than in the non-normal case because the model for the response variable given random effects is the normal distribution and that of the random

effects is the normal distribution as well. The two combined and integrating over the random effects leads to a normal distribution as the marginal model. In the non-normal case though, the model for the outcome variable and the random effects combined does not lead, in general, to closed-form solutions for the marginal model. Even if it does, expressions tend to be cumbersome. This is due to the lack of the elegant and convenient multivariate distributions analogous to the case of longitudinal data that can be assumed normally distributed. This poses computational and interpretational challenges. Specific to count data, which is of interest here, evaluation of the multivariate Poisson distribution grows in computational complexity with an increase in the dimensions due to the summations inherent in the distribution (Karlis, 2003). It is therefore of interest to find alternative means of analysis of correlated count data.

The generalized linear mixed model (GLMM) introduced in Section 3.3.4 and the combined model (Molenberghs *et al.*, 2007, 2010) introduced in Section 3.3.5 are alternatives. Some other references for interesting discussions on correlation and overdispersion are Winkelmann (2004, 2008); Sutradar (2011); Chin and Quddus (2003); Deb and Holmes (2000). The development of these two alternatives, and surrounding derivations have relevance well beyond mere data analysis. It so happens that the combined model can also be used to simulate correlated data. It is common practice in statistics to carry out Monte-Carlo (MC) simulations in which samples are randomly drawn from probability distributions to mimic statistical processes that can be used to study properties of statistical methods. Simulation of correlated Poisson random variables is a topic of ongoing research and various methods have been proposed in the literature to this end, some of which include: the overlapping sums (Madsen and Dalthorp, 2007; Mardia, 1970; Kocherlakota and Kocherlakota, 1992, 2001); Lognormal-Poisson hierachy; Normal to Anything (NorTA, Cario and Nelson, 1997, 1998; Nelsen, 2006; Mardia, 1970; Li and Hammond, 1975), and extensions thereof (Yahav and Shmueli, 2012; Ghosh and Pasupathy, 2012; Shin and Pasupathy, 2010; Avramidis *et al.*, 2009; Park and Shin, 1998; Downer and Moser, 2001). See also Devroye (1986) for an overview on random vari-

ate generation. These tools yield correlated Poisson random variables with the specification of the Poisson means and the desired or target correlation structure. Most of these methods, however, suffer from such limitations as: severe computational restrictions; difficulty achieving the target correlation; generated variables are required to be overdispersed; low correlations obtained; correlations constrained to be strictly positive; etc. Another approach is to use random effects to induce the correlation, thereby generating data from a hierarchical model. If the simulation is in the context of hierarchical models, this approach would be fine. However, whenever interest is in population-averaged or marginal models, the parameters used in the hierarchical model do not have a 1:1 correspondence with those in the marginal model. Given such a tool as the combined model that incorporates the two common features of count data, namely, overdispersion and correlation, it certainly is essential to generate data from such a method whenever interest is in simultaneously investigating these features. In this chapter, we present the combined model as a tool to generate correlated Poisson random variables.

6.2 Generation of Correlated Counts

Our focus in this chapter is the generation of correlated count or Poisson random variables for K independent subjects in a study with subject i having measurements Y_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$. This is based on specification of the mean model in terms of an $n_i \times p$ known design matrix X_i, a p-dimensional fixed-effects parameter vector $\boldsymbol{\xi}$ and $\boldsymbol{Z}_i$, an $n_i \times q$ design matrix for the random effects of subject i. From the combined model, if we assume the θ_{ij} in (3.16) to be independent as is often done in practice, then the association is only induced by the $\boldsymbol{b}_i$ and the θ_{ij} would cover the overdispersion not accounted for by the normal random effects. Then, Σ_i is reduced to a diagonal matrix. Alternatively, the θ_{ij} can be allowed to be correlated as well such that Σ_i can take on more general structures. This implies the use of some form of Multivariate Gamma (MGamma) distribution. For example, Σ_i can be chosen such that there is a time-dependence, or other covariate dependen-

cies, in the association structure. Evidently, as is also the case in the linear mixed model, when random effects and general Σ_i are present, the user needs to carefully ensure that the resulting marginal model is identifiable. A classical counterexample from the linear mixed model setting is a random intercept combined with a compound-symmetry residual structure. This leads to fully aliased parameters.

As will be presented in Section 6.2.1, the GLMM can be used to parsimoniously generate correlated count data with prespecified marginal mean function and such variance-covariance structures as compound symmetry and the one generated by random intercept and random slope. In the GLMM case, however, the random effects used do not separate correlation and overdispersion, a disadvantage that may lead to mis-representation of the random-effects variability. The algorithm for generating data from the combined model, which accounts for both correlation and overdispersion, is given in Section 6.2.2.

6.2.1 The GLMM as a Data Generator

The GLMM can be used to generate correlated random variables with a desired structure. Given a marginal mean $\boldsymbol{\mu}_i$, possibly depending on design matrix $\boldsymbol{X}_i^m$ of covariates, with superscript m indicating marginal, such that $\ln(\boldsymbol{\mu}_i) = \boldsymbol{X}_i^m\boldsymbol{\omega}$ where $\boldsymbol{\omega}$ are desired marginal parameters, and a variance-covariance matrix for $\boldsymbol{Y}_i$ denoted by V^m, correlated Poisson random variables can be generated from the GLMM using Algorithm 1 below;

Algorithm 1:

1. *Derive the unknowns $\boldsymbol{\xi}$ and D of the GLMM by comparing the desired marginals with the marginals from the GLMM.*
2. *Using D, simulate $\boldsymbol{b}_i$.*
3. *Compute* $\ln(\lambda_{ij}) = \boldsymbol{X}_{ij}^\top\boldsymbol{\xi} + \boldsymbol{Z}_{ij}^\top\boldsymbol{b}_i$.
4. *Simulate* $Y_{ij} \sim Poi(\lambda_{ij})$.

To put matters into context, if we consider the case of compound symmetry (CS), for example, in that the desired marginal mean is $\ln(\boldsymbol{\mu}_i) = \boldsymbol{X}_i^m\boldsymbol{\omega}$ and

desired variance-covariance structure is $V^m = M_i + \tau^2 J_i$ (CS structure), then the necessary unknowns in step 1 of the above algorithm are derived by comparing [a] $\boldsymbol{X}_i^m \omega = \boldsymbol{X}_i \boldsymbol{\xi} + 0.5 \boldsymbol{Z}_i D \boldsymbol{Z}_i^\top$ [which is (3.15a) expressed in matrix form] for the marginal mean, and, [b] $M_i + \tau^2 J_i = M_i + M_i(e^{\boldsymbol{Z}_i D \boldsymbol{Z}_i^\top} - J_i) M_i$ for the marginal variance-covariance structure. Solving [a] for $\boldsymbol{\xi}$ and [b] for D leads to:

$$\boldsymbol{\xi} = \left(X_i^\top X_i\right)^- X_i^\top (\boldsymbol{X}_i^m \omega - 0.5 \boldsymbol{Z}_i D \boldsymbol{Z}_i^\top), \tag{6.1a}$$

$$D = \left(\boldsymbol{Z}_i^\top \boldsymbol{Z}_i\right)^- \boldsymbol{Z}_i^\top \log\left(M_i^{-1} \tau^2 J_i M_i^{-1} + J_i\right) \boldsymbol{Z}_i \left(\boldsymbol{Z}_i^\top \boldsymbol{Z}_i\right)^-, \tag{6.1b}$$

where $(.)^-$ indicates a generalized inverse. For a general V^m, $\tau^2 J_i$ in D above becomes $V^m - M_i$. If the generalized inverse is not an inverse, the solution clearly is not unique. This is not a problem, it simply means that several choices of $\boldsymbol{\xi}$ and D are possible, that nevertheless all lead to the desired marginal structure. This is akin to the fact that there is a one-to-many map between a given marginal model on the one hand and the class of hierarchical models that marginalizes to it on the other. Any member of the class of hierarchical model can in principle be used as a data generator for the marginal structure.

6.2.2 The Combined Model as a Data Generator

The combined model, as well, can be used to generate correlated Poisson random variables following logic similar to that described in Section 6.2.1. The major difference from the GLMM is that there is a third unknown term in the combined model, i.e., Σ_i, the variance-covariance matrix for the overdispersion parameter(s). Given a desired mean and variance-covariance structure, Algorithm 2 generates the Poisson variates.

Algorithm 2:

1. *Derive the unknowns $\boldsymbol{\xi}$, D, and Σ_i in the CM.*
2. *Generate $\boldsymbol{\theta}_i \sim MGamma(mean = 1, variance = \Sigma_i)$.*
3. *Using D, simulate $\boldsymbol{b}_i$.*

4. *Compute* $\lambda_{ij}^* = \theta_{ij} \exp(x_{ij}^\top \boldsymbol{\xi} + z_{ij}^\top \boldsymbol{b}_i)$.
5. *Simulate* $Y_{ij} \sim Poi(\lambda_{ij}^*)$.

The necessary unknowns in step 1 of Algorithm 2 are given by $\boldsymbol{\xi}$ as in (6.1a) and further

$$D = \left(\boldsymbol{Z}_i^\top \boldsymbol{Z}_i\right)^- \boldsymbol{Z}_i^\top \log\left[M_i^{-1}(V^m - M_i)M_i^{-1} + J_i\right] \boldsymbol{Z}_i \left(\boldsymbol{Z}_i^\top \boldsymbol{Z}_i\right)^-,$$
$$\Sigma_i = e^{-\boldsymbol{Z}_i D \boldsymbol{Z}_i^\top}\left[M_i^{-1}(V^m - M_i)M_i^{-1} + J_i\right] - J_i,$$

where notational conventions are as before.

An extension to generating purely serially correlated outcomes may be achieved by removing the normal random effect and choosing $\boldsymbol{\theta}_i$ such that it follows a serially correlated multivariate gamma. Note that 'multivariate' is used here in the broad sense, because all hierarchical structures, such as longitudinal and clustered data to name a few, imply marginal multivariate structures. Evidently, in such structured designs, the marginal covariance matrix will typically not be unstructured.

The general form of the combined model (3.16), in the case of Poisson data, is that the normal random effects are correlated and the Gamma random effects are also correlated. From this general case, several special cases can be derived. An overview of the possible combinations is presented in Table 6.1.

The following special cases, which are also presented in Table 6.1, can be derived from the more general case:

- A combination of normal and independent Gamma random effects. This is the most commonly used form of the combined model in which the normal random effects induce/account for correlation while the Gamma random effects induce/account for overdispersion. It is model (3.16) but with Σ_i diagonal.

- Normal random effects without Gamma random effects. In this case, (3.16) reduces to (3.14) and data is generated as explained in Section 6.2.1. Here, the normal random effects induce/account for both correlation and overdispersion.

Table 6.1: *Possible combinations of the normal and Gamma random effects in the context of count data.* ✓ *refers to combinations of the combined model from which correlated and/or overdispersed data can be generated, while* ✗ *refers to the independent count data generation case*

			Gamma random effects		
	Present		Yes		No
			Correlated	Independent	
Normal random effects	Yes	Correlated	✓	✓	✓
		Independent	✓	✓	✓
	No		✓	✓	✗

- No normal random effects, no Gamma random effects. The absence of both random effects is equivalent to generating independent counts which is not of interest in this thesis.
- No normal random effects, correlated Gamma random effects such that both correlation and overdispersion are induced via the Gamma random effects. Thus, λ_{ij} in (3.16) becomes $\exp(\boldsymbol{X}_{ij}^{\top}\boldsymbol{\xi})$ and Σ_i is fully general.
- No normal random effects, independent Gamma random effects. In this case, the combined model reduces to the negative-Binomial model which accounts for overdispersion but not correlation. λ_{ij} in (3.16) becomes $\exp(\boldsymbol{X}_{ij}^{\top}\boldsymbol{\xi})$ and Σ_i is diagonal.

Extra variations can be constructed by choosing for the normal random effects (random intercept + slope, or higher dimensions) to be either independent (D diagonal) or correlated. In this chapter, we have only studied the latter case but the former is very easily obtainable.

6.3 Setup of Simulation Study

As illustrated in Section 6.2.2, the combined model can take on several forms or variations. To evaluate the performance of the different forms of the combined model as data generators, a simulation was set up across the variations. More specifically, given a pre-specified marginal mean and variance-covariance matrix, 1000 Monte Carlo replications of correlated count data sets were generated from each of the several forms. Marginal models were then fitted to these data sets and the difference between the pre-specified parameters and those estimated by fitting the marginal models were studied. Two different arms have been considered for the simulation, namely, sample size $K = 100$ and 500. For $K = 100$, 2 correlated Poisson variables were generated from the following model specification;

$$\begin{aligned} Y_{ij} &\sim \text{Poi}(\lambda^*_{ij}), \\ \lambda^*_{ij} = \theta_{ij}\lambda_{ij} &= \theta_{ij}\exp(\xi_0 + b_{0i} + \xi_1 T_i + (\xi_2 + b_{1i})t_{ij} + \xi_3 T_i * t_{ij}), \\ \boldsymbol{\theta}_i &\sim \text{MGamma}(\text{mean} = \mathbf{1}, \text{variance} = \Sigma_i), \\ \boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}\right], V^m = \begin{pmatrix} 36 & 12 \\ 12 & 29 \end{pmatrix}, \end{aligned} \tag{6.2}$$

where $T_i \sim$ Bernoulli(0.5), t_{ij} is the ordering of the j^{th} observation ($i = 1, \ldots, K = 100, j = 1, 2$) in subject i, and the desired marginal mean parameters are $\omega_0 = 1.521, \omega_1 = 0.237, \omega_2 = 0.254, \omega_3 = 0.345$. Generalized estimating equations (GEE1, Liang and Zeger 1986), NEGBIN, and the GLMM were used to study the behavior of the data generator, averaged over the 1000 MC replications. See Chapter 3 for a review of these methods.

For $K = 500$, 4 random variables were generated from a similar model as (6.2), the difference being that a random intercept model for the normal random effects was used. More specifically, the following specifications were

used;

$$\begin{aligned}\lambda_{ij}^{*} &= \theta_{ij}\lambda_{ij} = \theta_{ij}\exp(\xi_0 + b_{0i} + \xi_1 T_i + \xi_2 t_{ij} + \xi_3 T_i * t_{ij}),\\ \boldsymbol{b}_i &= b_{0i} \sim N(0, d),\\ V^m &= \begin{pmatrix} 256 & 128 & 144 & 224\\ 128 & 208 & 228 & 172\\ 144 & 228 & 299 & 296\\ 224 & 172 & 296 & 567 \end{pmatrix},\end{aligned} \tag{6.3}$$

where $i = 1, \ldots, K = 500$ and $j = 1, 2, 3, 4$. The desired marginal mean parameters were specified as $\omega_0 = 1.521, \omega_1 = 0.437, \omega_2 = -0.254$ and $\omega_3 = 0.145$. In addition to GEE1, NEGBIN and GLMM models used in the case of $K = 100$, the so-called marginal multilevel model (MMM) was also used, mainly motivated by the fact that the sensitivity of the MMM to starting values is less severe if the random intercept model is specified for the normal random effects than in the case of random intercept and slope. The MMM was described by Heagerty (1999) for binary longitudinal data, building on a specification of the marginal rather than the conditional mean given random effects. More precisely, this model puts together the two worlds of marginal and conditional or hierarchical modeling in the sense that it puts the ideas of GEE1 and the GLMM together leading to inferences both in the marginal and conditional senses.

6.4 Results of Simulation Study

Tables 6.2 and 6.6 present the results for the simulation study. Generally, from Table 6.2, all marginal models (GEE1, NEGBIN, MMM) seem to perform similarly across the various forms of the combined model. This is expected as the proposed data generator is aimed at the context of marginal models. Specific to this case of using a random-intercept model for the normal random effects, GEE1, MMM, and the GLMM yield the same results for time-related parameters ω_2, ω_3, ξ_2 and ξ_3 with minor differences between GEE1 or GLMM versus MMM in the case of normal and no gamma random effects. Given

Table 6.2: *Simulation, generate 4 random variables: Parameter estimates (standard deviations) for GEE1 (exchangeable correlation), NEGBIN, MMM and GLMM, and, absolute bias (MSE) for GEE1, NEGBIN and MMM, averaged over 1000 MC replications for sample size (K) = 500. True parameters are* $\omega_0 = 1.521, \omega_1 = 0.437, \omega_2 = -0.254$ *and* $\omega_3 = 0.145$ *and a random intercept model was specified for the normal random effects (RE). Corr means correlated while IND means independent.*

Model	Effect	Par.	Normal, Corr Gamma	Normal, IND Gamma	Normal, No Gamma	No normal, Corr Gamma	No normal, IND Gamma
			Parameter estimates (standard deviations)				
GEE1	intercept	ω_0	3.353 (0.098)	1.550 (0.515)	1.522 (0.046)	3.354 (0.098)	1.556 (0.509)
	T	ω_1	1.295 (0.103)	0.413 (0.564)	0.437 (0.057)	1.295 (0.103)	0.408 (0.561)
	t	ω_2	0.015 (0.040)	-0.296 (0.259)	-0.255 (0.018)	0.015 (0.040)	-0.298 (0.256)
	$t \cdot T$	ω_3	0.079 (0.041)	0.178 (0.281)	0.145 (0.022)	0.079 (0.041)	0.180 (0.279)
NEGBIN	intercept	ω_0	3.319 (0.185)	1.681 (0.633)	1.522 (0.046)	3.321 (0.185)	1.690 (0.629)
	T	ω_1	0.978 (0.193)	0.296 (0.678)	0.437 (0.057)	0.979 (0.194)	0.289 (0.677)
	t	ω_2	0.028 (0.085)	-0.354 (0.320)	-0.255 (0.018)	0.027 (0.085)	-0.358 (0.318)
	$t \cdot T$	ω_3	0.203 (0.088)	0.231 (0.341)	0.145 (0.022)	0.202 (0.088)	0.234 (0.340)
		γ	0.322 (0.009)	0.064 (0.004)	899.005 (1305.540)	0.322 (0.010)	0.064 (0.004)
MMM	intercept	ω_0	3.066 (0.118)	1.395 (0.617)	1.522 (0.050)	3.066 (0.118)	1.404 (0.608)
	T	ω_1	2.174 (0.142)	2.038 (0.669)	0.431 (0.062)	2.178 (0.141)	2.040 (0.664)
	t	ω_2	0.015 (0.040)	-0.296 (0.259)	-0.255 (0.022)	0.015 (0.040)	-0.298 (0.256)
	$t \cdot T$	ω_3	0.079 (0.041)	0.178 (0.281)	0.146 (0.022)	0.079 (0.041)	0.180 (0.279)
	d		1.448 (0.120)	5.630 (0.447)	0.006 (0.004)	1.451 (0.121)	5.660 (0.454)
GLMM	intercept	ξ_0	2.342 (0.136)	-1.421 (0.620)	1.519 (0.045)	2.341 (0.135)	-1.426 (0.612)

Continued on next page

Table 6.2 – continued from previous page

Model	Effect	Par.	Normal, Corr Gamma	Normal, IND Gamma	Normal, No Gamma	No normal, Corr Gamma	No normal, IND Gamma
	T	ξ_1	2.174 (0.142)	2.038 (0.669)	0.437 (0.055)	2.178 (0.141)	2.040 (0.664)
	t	ξ_2	0.015 (0.040)	-0.296 (0.259)	-0.255 (0.018)	0.015 (0.040)	-0.298 (0.256)
	$t \cdot T$	ξ_3	0.079 (0.041)	0.178 (0.281)	0.145 (0.022)	0.079 (0.041)	0.180 (0.279)
	d		1.448 (0.120)	5.630 (0.447)	0.006 (0.004)	1.451 (0.121)	5.660 (0.454)
				Absolute bias (MSE)			
GEE1	intercept	ω_0	1.832 (3.365)	0.029 (0.266)	0.001 (0.002)	1.833 (3.369)	0.035 (0.261)
	T	ω_1	0.858 (0.747)	0.024 (0.318)	0.000 (0.003)	0.858 (0.746)	0.029 (0.316)
	t	ω_2	0.269 (0.074)	0.042 (0.069)	0.001 (0.000)	0.269 (0.074)	0.044 (0.067)
	$t \cdot T$	ω_3	0.066 (0.006)	0.033 (0.080)	0.000 (0.000)	0.066 (0.006)	0.035 (0.079)
NEGBIN	intercept	ω_0	1.798 (3.268)	0.160 (0.426)	0.001 (0.002)	1.800 (3.273)	0.169 (0.424)
	T	ω_1	0.541 (0.331)	0.141 (0.479)	0.000 (0.003)	0.542 (0.331)	0.148 (0.481)
	t	ω_2	0.282 (0.087)	0.100 (0.113)	0.001 (0.000)	0.281 (0.086)	0.104 (0.112)
	$t \cdot T$	ω_3	0.058 (0.011)	0.086 (0.124)	0.000 (0.000)	0.057 (0.011)	0.089 (0.124)
MMM	intercept	ω_0	1.545 (2.401)	0.126 (0.396)	0.001 (0.002)	1.545 (2.402)	0.117 (0.383)
	T	ω_1	1.737 (3.037)	1.601 (3.011)	0.006 (0.004)	1.741 (3.051)	1.603 (3.008)
	t	ω_2	0.269 (0.074)	0.042 (0.069)	0.001 (0.000)	0.269 (0.074)	0.044 (0.067)
	$t \cdot T$	ω_3	0.066 (0.006)	0.033 (0.080)	0.001 (0.001)	0.066 (0.006)	0.035 (0.079)

normal random effects with random intercept only and no Gamma random effects, the marginal parameters are expected to be the same as the hierarchical parameters with a change in the intercept. Indeed, GEE1, NEGBIN, MMM, and GLMM yield the same parameter estimates with a change in the intercept (ξ_0) for GLMM. Across all variations of the combined model, GEE1, MMM, and GLMM generally differ on the intercept and treatment (T) parameters. No specific pattern can be identified for the NEGBIN relative to GEE1 and MMM, except in the above-mentioned case of normal random effects and no Gamma random effects. When the Gamma random effects are correlated, the parameter estimates are rather different from the true parameters and even change sign for ξ_2 and ω_2. Since the GLMM is a hierarchical model, the results for the GLMM presented should be interpreted with caution. We emphasize that GLMM should not be used to model data generated by our proposal.

From Table 6.6, which is the case of a random intercept and slope model for the normal random effects, both GEE1 and NEGBIN yield the same parameter estimates and standard deviations across the combined model variations. Since in this setting, only 2 random variables were generated, it may be interesting to consider the generation of more than 2 random variables and also larger sample sizes so as to get broader insight into this scenario. The parameter α for the NEGBIN goes to infinity in the absence of overdispersion, which is what we observe in the normal RE, no Gamma RE case. Again, the GLMM should be interpreted with care given that it is not a marginal but rather a hierarchical model.

Apart from the simulation, we also generated 4 different datasets of size $K = 500$ from the combined model with [1] two time points (bivariate case) with only the random intercept specified for the $\boldsymbol{b}_i$ random effects, [2] two time points with random intercept and slope, [3] four time points with random intercept only, and [4] four time points with random intercept and slope. The gamma random effects are correlated. Table 6.3 summarizes the generation settings considered here, in which 2 or 4 correlated Poisson variates are generated corresponding to 2 and 4 time points, respectively. We have only considered the case of the random intercept on the one hand and the random

Table 6.3: *Parameters specified to generate correlated Poisson random variables from the combined model.*

	2 time points	4 time points
	Case 1:	Case 3:
$\boldsymbol{X}_i^m = X_i$ covariates	*Intercept T t T*t*	*Intercept T t T*t*
$\boldsymbol{\omega}$	1.521 0.237 0.254 0.345	1.521 0.437 -0.254 0.145
$\boldsymbol{Z}_i$ covariates	*Intercept*	*Intercept*
V^m	$\begin{bmatrix} 36 & 12 \\ 12 & 29 \end{bmatrix}$	$\begin{bmatrix} 256 & 128 & 144 & 224 \\ 128 & 208 & 228 & 172 \\ 144 & 228 & 299 & 296 \\ 224 & 172 & 296 & 567 \end{bmatrix}$
	Case 2:	Case 4:
covariates ($\boldsymbol{X}_i^m = X_i$)	*Intercept T t T*t*	*Intercept T t T*t*
$\boldsymbol{\omega}$	2.521 0.237 0.254 0.345	1.521 0.437 -0.254 0.145
$\boldsymbol{Z}_i$ covariates	*Intercept + t*	*Intercept + t*
V^m	$\begin{bmatrix} 225 & 615 \\ 615 & 2581 \end{bmatrix}$	$\begin{bmatrix} 256 & 128 & 144 & 224 \\ 128 & 208 & 228 & 172 \\ 144 & 228 & 299 & 296 \\ 224 & 172 & 296 & 567 \end{bmatrix}$

intercept and slope in time models on the other, for illustrative purposes. It is easy to manipulate more general dimensions. Note though that the higher the random-effects dimension, the higher the risk of the D matrix not being positive-definite. Also, because the gamma random effects are allowed to be correlated, very little or no information is derived from the $\boldsymbol{b}_i$ random effects. We generate data given covariates ($\boldsymbol{X}_i^m$) as treatment (trt, 0 or 1), time (2 or 4 points) and the interaction of treatment and time. Note that we assume $\boldsymbol{X}_i^m = X_i$, thus using the same covariates but the method also allows for use

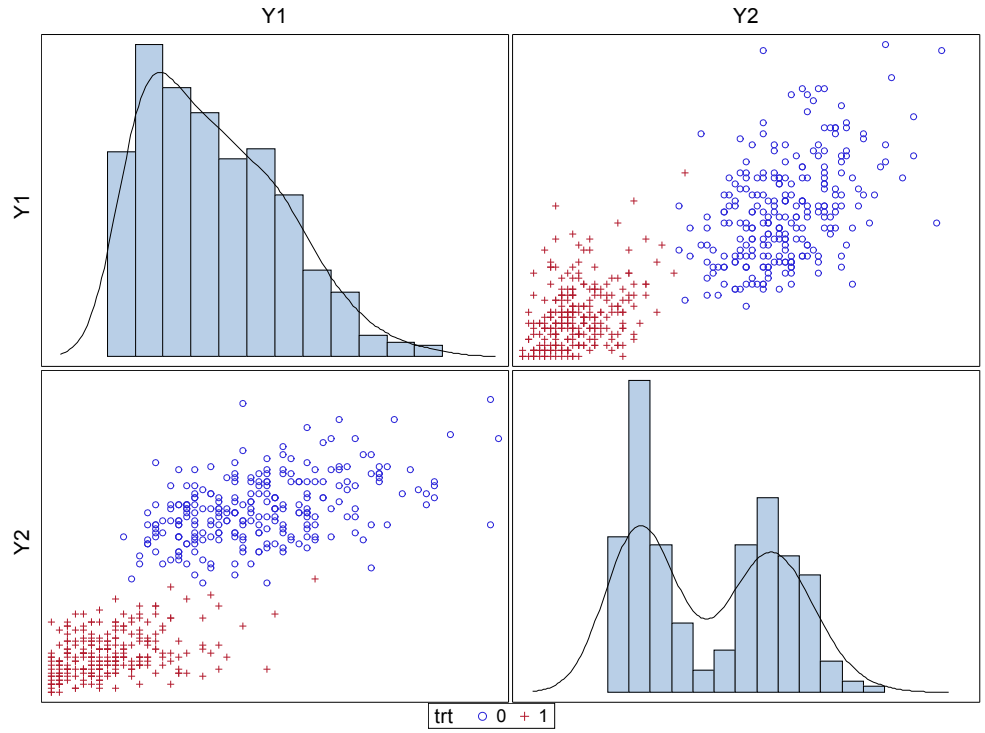


Figure 6.1: *Two Poisson random variables generated from the combined model with random intercept model.*

of different covariates in the two design matrices. Table 6.4 shows the results of the derived unknown parameters that aid the data generation process for the 4 cases presented in Table 6.3. Here, $\boldsymbol{\omega}$ is the parameter vector for the specified marginal mean and *diff* is the change between the marginal parameters $\boldsymbol{\alpha}$ and the conditional/derived parameters $\boldsymbol{\xi}$. As expected in the case of a random intercept model (cases 1 and 3), a change is only evident in the intercept relative to the other parameters. In cases 2 and 4 for the random intercept and slope model, a difference between the marginal and conditional mean parameters is reflected in the intercept and time parameter estimates. Table 6.5 presents the summary statistics and the Spearman correlation coefficients of the generated Poisson variables, while Figures 6.1–6.4 show marginal distributions and scatter plots of the generated random variables for cases 1–4, respectively. In Table 6.5, the mean is smaller than the square of the standard

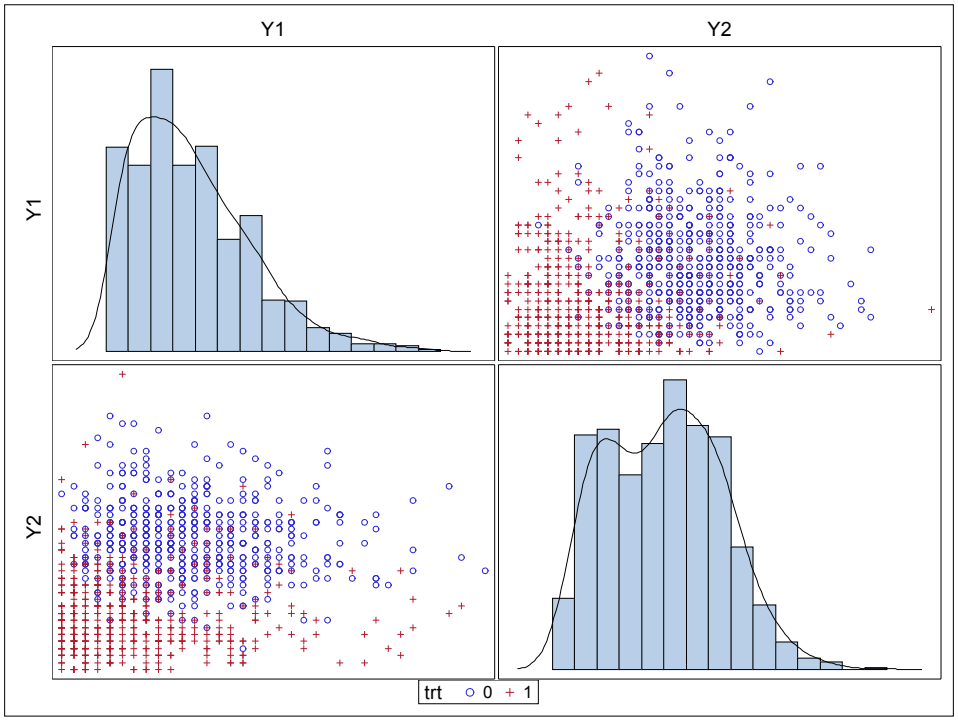


Figure 6.2: *Two Poisson random variables generated from the combined model with random intercept and slope model.*

deviation, indicating overdispersion. It can also be seen that the generated random variables are correlated (see ρ). From Table 6.3, cases 3 and 4 are similar with the only difference being that case 3 only has a random intercept while case 4 has random intercept and slope(time) as the covariates for the random effects. Specific to this case and given that Σ_i is fully general, there are minimal changes from case 3 to 4 (see Figures 6.3 and 6.4, and Table 6.5). Similarly, by comparing Figures 6.1 and 6.2, and also Table 6.5, we clearly see that that inclusion of a random slope allows to roughly retain the correlation structure, but modifies the mean and variance structures. Further, when the marginal structure is specified, it is possible to decompose the hierarchical structure (in particular, the random effects) in different ways, yet leading the same result, as it should be. Indeed, it is clear, from comparing Figures 6.3

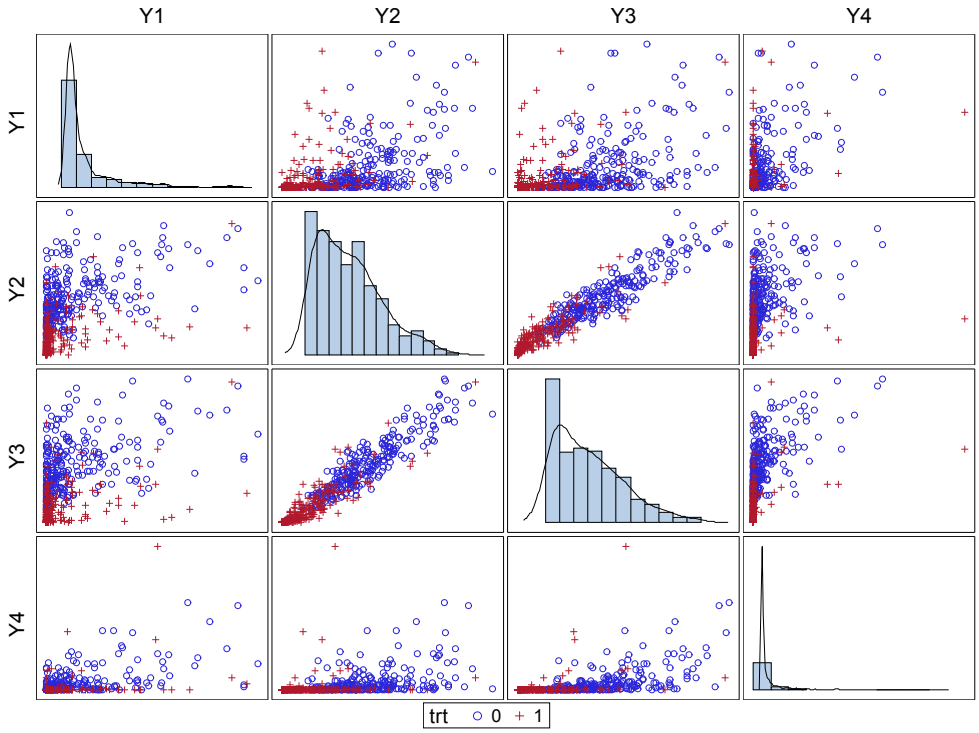


Figure 6.3: *Four Poisson random variables generated from the combined model with random intercept model.*

Table 6.4: *The necessary unknowns (ξ and D) for each of the cases presented in Table 6.3.*

		Derived unknowns			
Case	*Parameter*	$\boldsymbol{\omega}$	$\boldsymbol{\xi}$	*diff*	D
1.	Intercept	1.521	1.521	0.0002203	$\begin{bmatrix} 0.0004406 \end{bmatrix}$
	T	0.237	0.237	-1.02E-14	
	t	0.254	0.254	-6.88E-15	
	$t \cdot T$	0.345	0.345	1.082E-14	
2.	Intercept	2.521	2.521	-0.000493	$\begin{bmatrix} 0.000263 & -0.000039 \\ -0.000039 & 0.0006242 \end{bmatrix}$
	T	0.237	0.237	6.495E-15	
	t	0.254	0.253	0.0008976	
	$t \cdot T$	0.345	0.345	-3.89E-15	
3.	Intercept	1.521	1.518	0.002885	$\begin{bmatrix} 0.00577 \end{bmatrix}$
	T	0.437	0.437	-3.4E-14	
	t	-0.254	-0.254	-1.11E-15	
	$t \cdot T$	0.145	0.145	-1.5E-15	
4.	Intercept	1.521	1.520	0.0014135	$\begin{bmatrix} 0.0040014 & 0.0000601 \\ 0.0000601 & 0.0002349 \end{bmatrix}$
	T	0.437	0.437	-3.5E-14	
	t	-0.254	-0.255	0.0006473	
	$t \cdot T$	0.145	0.145	6.939E-16	

Table 6.5: *Summary statistics and the Spearman correlation (ρ) matrices of the generated Poisson variables; std refers to the standard deviation.*

Case	Var.	mean	std	median	min.	max.	ρ
1.	Y1	16.68	12.15	15.00	0	56	$\begin{bmatrix} 1 & 0.81 \\ & 1 \end{bmatrix}$
	Y2	28.36	19.93	27.50	0	75	
2.	Y1	87.07	43.92	85.50	7	205	$\begin{bmatrix} 1 & 0.88 \\ & 1 \end{bmatrix}$
	Y2	142.86	121.50	121.50	0	654	
3.	Y1	16.39	29.35	3	0	171	$\begin{bmatrix} 1 & 0.64 & 0.60 & 0.59 \\ & 1 & 0.96 & 0.69 \\ & & 1 & 0.79 \\ & & & 1 \end{bmatrix}$
	Y2	129.10	106.93	110	0	498	
	Y3	152.90	141.64	127	0	619	
	Y4	27.26	72.56	0	0	832	
4.	Y1	16.71	30.50	3.50	0	184	$\begin{bmatrix} 1 & 0.68 & 0.64 & 0.60 \\ & 1 & 0.96 & 0.68 \\ & & 1 & 0.79 \\ & & & 1 \end{bmatrix}$
	Y2	129.94	109.32	113.00	0	601	
	Y3	155.41	147.01	127.50	0	748	
	Y4	28.84	84.72	0	0	1147	

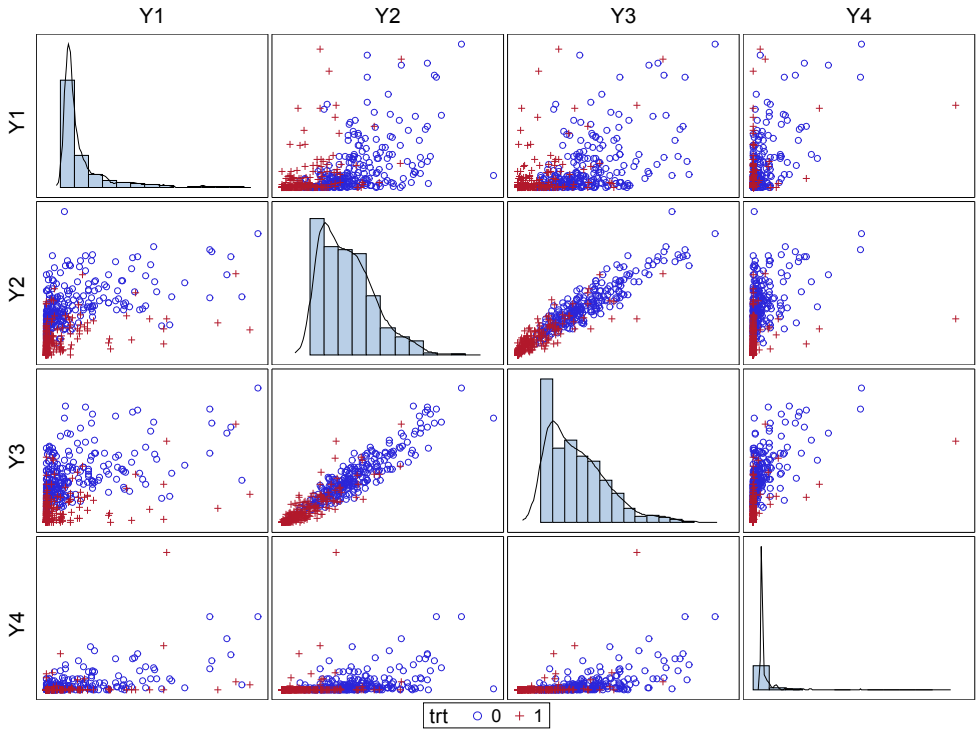


Figure 6.4: *Four Poisson random variables generated from the combined model with random intercept and slope model.*

Table 6.6: *Simulation, generate 2 random variables: Parameter estimates (standard deviations) for GEE1 (exchangeable correlation), NEGBIN and GLMM, and, absolute bias (MSE) for GEE1 and the NEGBIN models averaged over 1000 MC replications, N=100. True parameters are* $\omega_0 = 1.521, \omega_1 = 0.237, \omega_2 = 0.254, \omega_3 = 0.345$ *with a random intercept and slope model specified for the normal random effects (RE). Corr means correlated while IND means independent.*

Model	Effect	Par.	Normal, Corr Gamma	Normal, IND Gamma	Normal, No Gamma	No normal, Corr Gamma	No normal, IND Gamma
			Parameter estimates (standard deviations)				
GEE1	intercept	ω_0	1.749 (0.240)	1.522 (0.304)	1.524 (0.127)	1.752 (0.238)	1.497 (0.304)
	T	ω_1	0.922 (0.258)	0.226 (0.350)	0.241 (0.157)	0.866 (0.260)	0.253 (0.351)
	t	ω_2	0.266 (0.127)	0.249 (0.175)	0.252 (0.078)	0.263 (0.126)	0.264 (0.175)
	$t \cdot T$	ω_3	0.336 (0.136)	0.355 (0.198)	0.342 (0.095)	0.337 (0.137)	0.338 (0.199)
NEGBIN	intercept	ω_0	1.749 (0.240)	1.522 (0.304)	1.524 (0.127)	1.752 (0.238)	1.497 (0.304)
	T	ω_1	0.922 (0.258)	0.226 (0.350)	0.241 (0.156)	0.866 (0.260)	0.253 (0.351)
	t	ω_2	0.266 (0.127)	0.249 (0.175)	0.253 (0.078)	0.263 (0.126)	0.264 (0.175)
	$t \cdot T$	ω_3	0.336 (0.136)	0.355 (0.198)	0.342 (0.095)	0.337 (0.137)	0.338 (0.199)
		γ	6.494 (1.628)	3.715 (0.729)	1238.877 (1640.875)	6.316 (1.492)	3.740 (0.714)
GLMM	intercept	ξ_0	1.373 (0.244)	1.003 (0.311)	1.508 (0.066)	1.373 (0.240)	0.988 (0.311)
	T	ξ_1	1.186 (0.263)	0.484 (0.353)	0.218 (0.082)	1.129 (0.261)	0.502 (0.353)

Continued on next page

Table 6.6 – continued from previous page

Model	Effect	Par.	Normal, Corr Gamma	Normal, IND Gamma	Normal, No Gamma	No normal, Corr Gamma	No normal, IND Gamma
	t	ξ_2	0.427 (0.127)	0.464 (0.177)	0.242 (0.044)	0.425 (0.126)	0.474 (0.176)
	$t \cdot T$	ξ_3	0.223 (0.136)	0.252 (0.198)	0.383 (0.057)	0.225 (0.136)	0.240 (0.199)
	d_{11}		0.806 (0.224)	1.997 (0.457)	0.043 (0.069)	0.824 (0.219)	1.980 (0.446)
	d_{12}		-0.339 (0.105)	-1.058 (0.255)	-0.031 (0.039)	-0.347 (0.103)	-1.051 (0.246)
	d_{22}		0.155 (0.052)	0.590 (0.147)	0.225 (0.141)	0.159 (0.052)	0.588 (0.140)
				Absolute bias (MSE)			
GEE1	intercept	ω_0	0.228 (0.110)	0.001 (0.093)	0.003 (0.016)	0.231 (0.110)	0.024 (0.093)
	T	ω_1	0.685 (0.536)	0.011 (0.123)	0.004 (0.025)	0.629 (0.463)	0.016 (0.124)
	t	ω_2	0.012 (0.016)	0.005 (0.031)	0.002 (0.006)	0.009 (0.016)	0.010 (0.031)
	$t \cdot T$	ω_3	0.009 (0.018)	0.010 (0.039)	0.003 (0.009)	0.008 (0.019)	0.007 (0.040)
NEGBIN	intercept	ω_0	0.228 (0.110)	0.001 (0.093)	0.003 (0.016)	0.231 (0.110)	0.024 (0.093)
	T	ω_1	0.685 (0.536)	0.011 (0.123)	0.004 (0.025)	0.629 (0.463)	0.016 (0.124)
	t	ω_2	0.012 (0.016)	0.005 (0.031)	0.001 (0.006)	0.009 (0.016)	0.010 (0.031)
	$t \cdot T$	ω_3	0.009 (0.018)	0.010 (0.039)	0.003 (0.009)	0.008 (0.019)	0.007 (0.040)

and 6.4, that the same marginal structure (mean, variance, correlation) can be obtained, with our without the use of a random slope. This gives the user some latitude as to choose a decomposition that is flexible yet computationally efficient.

6.5 Discussion and Conclusions

The combined model as introduced by Molenberghs *et al.* (2007, 2010) simultaneously accommodates correlation and overdispersion unexplained by the normal random effects. In the absence of correlation, the model simplifies to a negative-binomial model for overdispersion. On the other hand, in the absence of overdispersion, it simplifies to the GLMM. The model's flexible capabilities make it a good candidate as a data generator given that one always wants to generate data that reflects the characteristics of interest, in this case, overdispersion and/or correlation. The CM is a convenient tool that mimics or incorporates these intrinsic features of correlated count data. In particular, a fully marginal view as well as a random-effects view can be taken. This implies that a broad toolkit emerges. In the purely marginal view, essentially a multivariate gamma variate, easy to generate, is transformed to a multivariate count variable.

The covariates determining the fixed- and random-effects design matrices are kept simple herein. This is not limiting in the sense that a specification of any covariates can be done as is needed. It is possible to encounter non-positive definite D matrices or negative entries along the diagonal of Σ_i. This may point to a non-allowable hierarchical model to come with the marginal model or perhaps a marginal model that is in itself not allowable. The analogy would be a multivariate normal with a given but non-positive definite variance-covariance matrix. Such model is invalid in the first place and needs to be reconsidered.

Because the combined model is hierarchical, random variables with only positive correlations are generated due to restrictions of positive-definiteness on the random effects variance-covariance matrices. This may be a drawback

for the combined model, as is the case for some of the methods present in literature for count data generation. However, a way to overcome this is to generate directly from the marginal model, arguably via correlated θ_{ij}, of which the variance-covariance matrix Σ_i then reflects the desired structure.

Part III

Software Contributions

Chapter 7

Pairwise Likelihood as a Marginal Model Approach to Hierarchical Count Data using SAS Software

7.1 Introduction

Models for data arising from counting have been quite extensively studied in literature. They have also been widely implemented in various software packages like SAS, R, Matlab, SPPS, and so on, for various contexts as overdispersion, correlation and zero-inflation, with several extensions. For example, Zeileis *et al.* (2008) gives an overview of regression models for count data in R. See also Cameron and Trivedi (2013) and the references therein.

Extensions of count data collected repeatedly over time for the same subject are also commonly encountered in scientific research. Contemporary studies frequently aim at describing the evolution of subjects over time or observing more than one response from a single subject. The features of overdispersion and zero-inflation carry over in the correlated case and have also been studied quite in detail. For example, Breslow and Clayton (1993) and Wolfinger and O'Connell (1993) extended the generalized linear modeling (GLM) framework to the so-called generalized linear mixed model (GLMM) in which the correla-

tion is accounted for by use of random effects. Molenberghs *et al.* (2007, 2010) propose a joint model for clustering and over-dispersion through two separate sets of random effects.

In the maximum likelihood framework, the multivariate distribution is used to model correlated data. For continuous data, calculations are feasible because of the closed form expressions for the marginal distribution. This has the advantage of a gain in efficiency as long as the model is correctly specified. However, use of a so-called Multivariate Poisson (MP) model is constrained by the complexity of the probability function to be calculated. This is because it involves summations which may increase the computational burden with increase in the number of measurements per subject and/or sample size. See, for example, Karlis (2003), Karlis and Ntzoufras (2003), Kocherlakota and Kocherlakota (2001), among others, for details. In this Chapter, we present a SAS macro called **PLCounts** that uses the method of pseudo-likelihood, taking the form of pairwise likelihood, developed in Chapter 4. Pseudo-likelihood has the advantage of drastically simplifying computation while retaining sufficiently high statistical efficiency, but, also allows in this case for inference not only on the marginal mean parameters but also the covariance structure. Because the most common tool used for analysis of correlated count data for inferences on the marginal mean is GEE (Liang and Zeger, 1986), we compare pairwise likelihood to GEE by analyzing the Epilepsy and Whitefly datasets introduced in Sections 2.2 and 2.5, respectively. Section 7.2 introduces the SAS macro and demonstrates how to call the macro to analyze correlated count data plus the output of the macro.

7.2 The SAS Macro

7.2.1 Introduction

We have implemented the pseudo-likelihood approach presented in Section 4.2 in a SAS macro which we have named **PLCounts**. The macro was developed and tested in SAS version 9.1.3 (SAS Institute Inc., 2002-2004) although it should execute in other SAS versions without problems. Ta-

ble 7.1 presents the arguments that **PLCounts** uses to facilitate the fitting of pairwise likelihood to correlated count data. **PLCounts** begins with some pre-processing of the data, obtaining initial values for the marginal mean parameters by fitting a univariate Poisson regression using the **GENMOD** procedure and the creation of the design matrix using the SAS **LOGISTIC** procedure, even in the presence of classification variables. By making use of these procedures, any parameterization method possible in the SAS **LOGISTIC** and **GENMOD** procedures can be used for the design matrix, e.g., effect, glm, ordinal, reference, etc. We refer to the SAS website (`http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect006.htm`) for documentation about the different parameterization methods for classification variables. Estimation of the parameter $\boldsymbol{\lambda} = (\boldsymbol{\beta}, \theta_{ist})^T$ is done in SAS/IML using the Newton-Raphson (NR) algorithm. Appendix A.2 shows the gradient and Hessian functions of the log PL function in (3.27), with respect to the unknown parameters in $\boldsymbol{\lambda}$, which are supplied to the NR optimization step.

Table 7.1: *The macro arguments for* **PLCounts** *and their corresponding description.*

Macro argument	**Description**
`DATA`	Dataset containing subject or cluster identification variable and covariates or fixed effects from which design matrix X_i is created. This is a required argument. This dataset should take the "long" or hierarchical data structure as opposed to the wide format. Please note that all variables to be specified in the other macro arguments must be in this dataset.
	Continued on next page

Table 7.1 – continued from previous page

Macro argument	Description
SUBJECT	Subject identification variable in DATA
RESPONSE	Name of the response variable; must be in the dataset specified with the DATA argument
TIMEVAR	Variable with ordering of the observations within subject. This should be in DATA dataset.
FIXED	Covariates for the marginal mean from which X_i is created e.g., FIXED = trt time trt*time. Please note that the intercept must not be included in the specification of FIXED. It is added at the creation of the X_i design matrix.
CLASS	Specify all classification variables included in the XCOV argument, for example, CLASS = trt.
INIT_THETA0	Initial value for the covariance parameter θ_0. By default, INIT_THETA0 = 0.3.
CORR_DATASET	Name to give to dataset containing the correlations of all the subjects for all the timepoints. If left blank, no correlations are calculated. Default is blank.
outPL	Name of final output dataset containing the parameter estimates and standard errors from the pairwise approach. By default, OUTPL = solutionpseudo.

Continued on next page

Table 7.1 – continued from previous page

Macro argument	Description
PARAM	Parameterization method for the classification variables specified in the CLASS argument. Default is PARAM = glm. See SAS documentation for details about the different methods at http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect006.htm.
ABSGTOL	Specifies the absolute gradient stopping criterion. By default, ABSGTOL = 0.00001.
EXP_COV	To keep the covariance (θ_0) strictly positive by exponentiation, set EXP_COV = 1. By default, EXP_COV = 0 so that θ_0 can take on either positive or negative values.

7.2.2 Analyzing the Epilepsy Dataset

We analyze data on Epilepsy introduced in Section 2.2 using **PLCounts** and compare results from fitting a univariate Poisson regression, GEE and pairwise likelihood. The following code calls macro **PLCounts** to model the evolution of the number of epileptic seizures (nseizw) between the two treatment (trt) arms:

```
%PLCounts(DATA = epi, SUBJECT = id, CLASS = trt,
RESPONSE = nseizw, FIXED = trt studyweek studyweek*trt,
INIT_THETA0 = 0.3, TIMEVAR = studyweek, CORR_DATASET = corr,
EXP_COV = 0);
```

Please note that not all the macro arguments are presented in the above

macro call. However, all the arguments used for macro **PLCounts** are defined in Table 7.1. The top panel of Figure 7.1 shows the output as printed to the output window while a dataset called *solutionpseudo* (`outPL = solutionpseudo` by default) containing the parameter estimates, standard errors, 95% confidence intervals and the p-values is created. Also output are the results of the univariate Poisson regression in a dataset called *solutionuni*.

To correct for baseline characteristics of the patients, namely, race, age, gender, height, and weight, in which race and gender (sex) are classification variables, the following code is invoked:

```
%PLCounts(DATA = epi, SUBJECT = id, CLASS = trt race sex,
FIXED = trt studyweek studyweek*trt race age sex height weight,
INIT_THETA0 = 0.3, TIMEVAR = studyweek, CORR_DATASET = corr,
RESPONSE = nseizw, EXP_COV = 0);
```

Note that the only difference between the two macro calls above is that variables `race` and `sex` have been added to the `class` argument and variables `race`, `age`, `sex`, `height`, `weight` to the `FIXED` argument. The bottom panel of Figure 7.1 shows the output resulting from the inclusion of patient characteristics in the model as printed to the output window while Table 4.5 shows the parameter estimates and standard errors for the univariate, GEE and pairwise likelihood approaches. Similar results are observed for GEE and PL, especially as far as the standard errors are concerned. Parameter estimates and standard errors in Figure 7.1 are different from those of the pseudo-likelihood part of Table 4.5 because in Figure 7.1, the placebo group (trt=0) is the reference category while in Table 4.5, the AED group (trt=1) is the reference category. Note also that in Table 4.5, the reference categories are dropped.

7.2.3 Analyzing the Whitefly Dataset

The whitefly data introduced in Section 2.5 is analyzed by modeling the evolution of the number of immature whiteflies (`imm`) as a function of `block` for the treatment groups (`trt`) by invoking the following code:

```
%PLCounts(DATA = Whitefly, SUBJECT = plantid, RESPONSE = imm,
```

Obs	Parameter	Level1	Estimate	Robust standard error	95% Lower Confidence Limit	95% Upper Confidence Limit	Chisquare value	P-value
1	Intercept		0.8502	0.2443	0.3714	1.3289	12.11	0.0005
2	trt	AED	0.0642	0.4142	-0.7477	0.8761	0.02	0.8768
3	trt	Placebo	0.0000	0.0000	0.0000	0.0000	.	.
4	studyweek		-0.0105	0.0303	-0.0698	0.0488	0.12	0.7295
5	studyweek*trt	AED	-0.0284	0.0356	-0.0982	0.0413	0.64	0.4240
6	studyweek*trt	Placebo	0.0000	0.0000	0.0000	0.0000	.	.
7	Theta0		1.1017	0.2699	0.5726	1.6308	16.66	<.0001

Obs	Parameter	Level1	Estimate	Robust standard error	95% Lower Confidence Limit	95% Upper Confidence Limit	Chisquare value	P-value
1	Intercept		3.8376	5.0362	-6.0333	13.7084	0.58	0.4461
2	trt	AED	0.0805	0.4033	-0.7101	0.8710	0.04	0.8419
3	trt	Placebo	0.0000	0.0000	0.0000	0.0000	.	.
4	studyweek		-0.0116	0.0312	-0.0728	0.0497	0.14	0.7113
5	studyweek*trt	AED	-0.0225	0.0330	-0.0871	0.0422	0.46	0.4955
6	studyweek*trt	Placebo	0.0000	0.0000	0.0000	0.0000	.	.
7	race	1	-0.0774	0.5479	-1.1512	0.9964	0.02	0.8876
8	race	2	0.0000	0.0000	0.0000	0.0000	.	.
9	age		-0.0203	0.0194	-0.0582	0.0177	1.10	0.2954
10	sex	1	0.7755	0.4402	-0.0872	1.6382	3.10	0.0781
11	sex	2	0.0000	0.0000	0.0000	0.0000	.	.
12	height		-0.0362	0.0711	-0.1755	0.1031	0.26	0.6109
13	weight		-0.0024	0.0083	-0.0186	0.0139	0.08	0.7770
14	Theta0		1.0793	0.2515	0.5864	1.5723	18.41	<.0001

Figure 7.1: *Epilepsy Data. Results output by macro* **PLCounts** *to the SAS output window. Top panel is without patients characteristics while bottom panel corrects for patient characteristics. Please note that the two panels are output by two separate calls of* **PLCounts**.

Table 7.2: *Whitefly data: Parameter estimates (standard errors) for a univariate Poisson model, GEE (exchangeable correlation) and pseudo-likelihood (3.27).*

Parameter	Univariate	GEE	Pseudo-likelihood
Intercept	1.1405 (0.0578)	1.1638 (0.1877)	1.1434 (0.1842)
block(1)	0.0270 (0.0402)	0.0197 (0.1418)	0.0245 (0.1428)
block(2)	0.1535 (0.0390)	0.1432 (0.0803)	0.1517 (0.0804)
trt(1)	-1.0642 (0.0762)	-1.0586 (0.1384)	-1.0560 (0.1394)
trt(2)	-1.3630 (0.0858)	-1.3649 (0.2525)	-1.3519 (0.2424)
trt(3)	-2.0746 (0.1169)	-2.0823 (0.2756)	-2.0125 (0.2724)
trt(4)	-1.7587 (0.1005)	-1.7538 (0.1601)	-1.7415 (0.1593)
trt(5)	1.3533 (0.0431)	1.3281 (0.1202)	1.3561 (0.1206)
week	0.0902 (0.0048)	0.0901 (0.0213)	0.0894 (0.0213)
θ_{st}			0.0000 (0.0279)

```
CLASS = plantid trt block, FIXED = block trt week,
INIT_THETA0 = 2, TIMEVAR = week, EXP_COV = 1);
```

The results are shown in Table 7.2 in comparison to the univariate Poisson regression and GEE. Note that `EXP_COV = 1` meaning that the covariance (θ_{st}) is constrained to be strictly positive. Though θ_{st} hails from a Poisson distribution and is expected to be strictly positive, this interpretation takes effect in a hierarchical modeling framework. In the context of marginal models, this parameter can also take on negative values. This phenomenon is often a source of confusion, and it is less well understood in non-Gaussian cases than for continuously distributed hierarchical data. Pryseley *et al.* (2011) describe how such negative correlations can be estimated and interpreted for both Gaussian and non-Gaussian settings. One important situation where negative association is natural is where cluster members are in a competitive relation with one another. Molenberghs and Verbeke (2011) further discuss how a negative

correlation can be reconciled with a hierarhical model interpretation.

7.3 Concluding Remarks

Pseudo-likelihood (PL), or more specifically, pairwise likelihood is a viable alternative to generalized estimating equations (GEE) when modelling different data types, including but not limited to, correlated count data. PL, like GEE, yields consistent and asymptotically normally distributed parameter estimates with a sandwich estimator used to calculate the variance. On the one hand, GEE remains computationally faster than PL because it only evaluates the first moment and plugs in working assumptions for the second. On the other, it allows for the misspecification of the working correlation structure implying that one cannot rely on the correlation estimates from GEE for formulating answers to scientific questions, should interest be in the association as well.

Our pairwise approach can be used when scientific interest is not only in the marginal mean parameters but also in the association between pairs of measurements, unlike GEE. The method of pseudo-likelihood has been studied for such cases as, for example, binary data and its applications are prevalent in the literature. We were, however, not aware of its application in the context of correlated count data and have therefore developed a SAS macro called **PLCounts** to fill the void in the context of marginal models. Of course, one could consider a very general correlation structure with GEE, but this cannot be subjected to standard statistical assessment, e.g., hypothesis-testing based assessment. Alternatively, one could switch to second-order GEE (Zhao and Prentice, 1990), but this may come with considerable computational complexity in which case, pairwise likelihood becomes a potential candidate. Note though that the number of measurements per subject or cluster size (n_i) is a determinant of the computational burden in the PL approach presented herein because evaluation of the marginal PL is done for all $[n_i(n_i - 1)]/2$ possible pairs of a subject.

Our method assumes the covariance to be exchangeable, meaning that it is assumed the same for all the pairs and subjects. This assumption can

however be relaxed and will be implemented in subsequent versions of the macro available at `http://ibiostat.be/software/count`.

Chapter 8

%GEE2Counts: A SAS Macro for Modeling Correlated Counts using Second-order Generalized Estimating Equations

8.1 Introduction

The estimating equations presented in Chapter 5 have been implemented in a SAS macro called **GEE2Counts** that is described in Section 8.2.1. We demonstrate the functionality of the macro by analyzing the Jimma dataset introduced in Sections 2.3 and show the results as output by the macro.

8.2 The SAS Macro

8.2.1 Introduction

Macro **GEE2Counts** was developed and tested in SAS version 9.3 although it should execute in other SAS versions without problems. The results in this chapter were obtained from SAS version 9.4. In Table 8.1, the arguments

that **GEE2Counts** uses to facilitate the fitting of the estimating equations to correlated count data are presented. **GEE2Counts** also begins with some preprocessing of the data and the partial creation of the design matrix (5.8) using the **LOGISTIC** procedure whether or not there be classification variables. Initial values for the marginal mean parameters $\boldsymbol{\xi} = (\xi_{0s}, \xi_{0t}, \xi_{0st}, \xi_1, \xi_2, \ldots, \xi_p)^\top$ are obtained by fitting a univariate Poisson regression using the **GENMOD** procedure, except for the intercepts ξ_{0s}, ξ_{0t} and ξ_{0st} which are set equal to the value of the intercept from the univariate Poisson regression by default. Estimation of the parameters $\boldsymbol{\xi}$ is done in SAS/IML.

Table 8.1: *The macro arguments for* **GEE2Counts** *and their corresponding description.*

Macro argument	**Description**
`DATA`	Dataset containing subject or cluster identification variable and covariates or fixed effects from which design matrix is created. This is a required argument. This dataset should take the "long" or hierarchical data structure as opposed to the wide format. Please note that all variables to be specified in the other macro arguments must be in this dataset.
`ID`	Subject identification variable in `DATA`.
`RESPONSE`	Name of the response variable; must be in the dataset specified with the `DATA` argument.
`TIMEVAR`	Variable with ordering of the observations within subject. This should be in `DATA` dataset.

Continued on next page

Table 8.1 – continued from previous page

Macro argument	Description
`FIXEDX`	Covariates for the marginal mean from which $\boldsymbol{X}_{is}$ and $\boldsymbol{X}_{it}$ are created e.g., `FIXEDX = trt time trt*time`. Please note that the intercept must not be included in the specification of `FIXEDX` as it is added internally.
`CLASSX`	Specify all classification variables included in the `FIXEDX` argument, for example, `CLASSX = trt`.
`FIXEDCOV`	Covariates for the marginal mean from which $\boldsymbol{X}_{ist}$ is created e.g., `FIXEDCOV = gender age`. Again, the intercept must not be included in the specification of `FIXEDCOV` as it is added internally.
`CLASSCOV`	Specify all classification variables included in the `FIXEDCOV` argument, for example, `CLASSX = gender`.
`PARAM`	Parameterization method for the classification variables specified in the `CLASSX` and `CLASSCOV` arguments. Default is `PARAM = glm`. See SAS documentation for details about the different methods at `http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect006.htm`.

Continued on next page

Table 8.1 – continued from previous page

Macro argument	Description
`TIMEVARYFUNC`	Specify the time-varying function to be used for the creation of $\boldsymbol{X}_{ist}$. By default, `TIMEVARYFUNC = 0` meaning that there are no time-varying covariates specified in `FIXEDCOV`. `TIMEVARYFUNC = 1` would resulting in using the difference in the time-varying covariates between time point s and t. Other possibilities are 2 for the ratio, 3 for the product, 4 for the sum and 5 for the lag.
`TIMEVARYING`	The time-varying covariates are specified using this argument. It is blank by default meaning that, again, a time-stationary model would be used for the covariance.
`OUT`	Specify the name of the output dataset for the final parameters $\boldsymbol{\xi}$ and their corresponding standard errors.
`MAXITER`	Specify the maximum number of iterations, by default equal to 200.

Continued on next page

Table 8.1 – continued from previous page

Macro argument	Description
`RANDINTERCEPT`	Specify the initial values for the 3 intercepts ξ_{0s}, ξ_{0t} and ξ_{0st}. By default, `RANDINTERCEPT = 0` such that these intercepts are all set equal to the intercept from the univariate Poisson regression. Setting `RANDINTERCEPT = 1` results in randomly generating these 3 values from a normal distribution with zero mean. It is also possible for the user to specify these values by using the `INIT` argument, e.g., `INIT = -0.6 -1.2 1.25`. In our experience, while the default approach worked fine in some cases, there were others that were affected by these starting values which motivated the random generation and the user specification of the values. If both `INIT` and `RANDINTERCEPT` are specified, `INIT` takes precedence.
`EPSILON`	Specifies the stopping criterion. By default, `EPSILON = 0.00001`.
`CORR_DSN`	Name to give to dataset containing the correlations of all the subjects for all the time points. If left blank, no correlations are calculated. It is set to `CORR_DSN = corr` by default.

8.2.2 Analyzing the Epilepsy Dataset

The following code fits the time-stationary model (5.9):

```
%GEE2Counts(data = epi59, id = id, timevar = time,
classX = trt, param = glm, response = counts,
```

```
fixedX = trt|time baseline age, classCov = trt,
fixedCov = trt age, timevaryfunc = 0, timevarying =,
out = final, epsilon = 0.00001, weights = 0, randIntercept = 0
init = 0.1055 -0.5665 2.2977);
```

while (5.10) is fitted by setting `fixedCov = time trt age`, `timevaryfunc = 1`, and `timevarying = time`. The results of these model fits are shown in Figures 8.1 and 8.2. Note that both the model based and empirical standard errors, their corresponding χ^2 values and p-values are reported in the output.

8.3 Concluding Remarks

We have presented a SAS macro that can be used to model correlated count data with the intention of drawing inference on the marginal mean parameters as well as the association structure. The macro automatically handles the creation of the design matrix such that one need not worry about manually creating dummy variables in the presence of classification variables. The user however may need to be careful about the initial values of the 3 intercepts ξ_{0s}, ξ_{0t} and ξ_{0st} since, in our experience, computational issues sometimes resulted and would be resolved by simply changing these initial values. Changing them can be done by either randomly generating them from a normal distribution with zero mean or by the user specifying the values.

```
            GEE2 applied to Longitudinal Count Data with
                    time stationary covariates

                  Number of Clusters: 59
                Number of observations: 236
                 Minimum Cluster Size: 4
                 Maximum Cluster Size: 4
                 Number of Iterations: 15
                      Outcome Variable: counts
       Covariates used for [X_is,X_it]: trt|time baseline age
       Covariates used for [X_ist] or Cov(Y_is,Yit): trt age

                 Marginal Mean Parameter Estimates
              and Naive (Model-Based) Standard Errors

Parameter Level1    Initial value Estimate   StdErr    ChiSq Pvalue

int_s                      0.1055  -1.1904   0.2157    30.45 <.0001
int_t                     -0.5665  -1.2119   0.2474    24.00 <.0001
int_st                     2.2977   1.3148   0.0929   200.17 <.0001
time                      -0.0730  -0.1724   0.0396    18.95 <.0001
trt       Placebo          0.1165   0.2782   0.0452    37.89 <.0001
trt       Progabide        0.0000   0.0000   0.0000        .     .
time*trt  Placebo          0.0315   0.0404   0.0390     1.07 0.3004
time*trt  Progabide        0.0000   0.0000   0.0000        .     .
baseline                   0.0225   0.0379   0.0015   629.29 <.0001
age                        0.0163   0.0082   0.0027     8.93 0.0028

                  Marginal Mean Parameter Estimates
                     and Sandwich Standard Errors

Parameter Level1    Initial value Estimate   StdErr    ChiSq Pvalue

int_s                      0.1055  -1.1904   1.2655     0.88 0.3468
int_t                     -0.5665  -1.2119   1.2563     0.93 0.3347
int_st                     2.2977   1.3148   0.4836     7.39 0.0066
time                      -0.0730  -0.1724   0.0500    11.87 0.0006
trt       Placebo          0.1165   0.2782   0.2490     1.25 0.2638
trt       Progabide        0.0000   0.0000   0.0000        .     .
time*trt  Placebo          0.0315   0.0404   0.1576     0.07 0.7977
time*trt  Progabide        0.0000   0.0000   0.0000        .     .
baseline                   0.0225   0.0379   0.0088    18.46 <.0001
age                        0.0163   0.0082   0.0129     0.40 0.5257
```

Figure 8.1: *Epilepsy Data: Results as output by macro* **GEE2Counts** *after fitting (5.9).*

```
GEE2 applied to Longitudinal Count Data with time varying covariates
 (difference function applied to time varying covariate(s) [time])

                    Number of Clusters: 59
                  Number of observations: 236
                   Minimum Cluster Size: 4
                   Maximum Cluster Size: 4
                   Number of Iterations: 15
                        Outcome Variable: counts
          Covariates used for [X_is,X_it]: trt|time baseline age
        Covariates used for [X_ist] or Cov(Y_is,Yit): time trt age

                  Marginal Mean Parameter Estimates
                and Naive (Model-Based) Standard Errors

  Parameter Level1    Initial value Estimate   StdErr    ChiSq Pvalue

  int_s                      0.1055  -1.3907   0.2221    39.20 <.0001
  int_t                     -0.5665  -1.6024   0.2518    40.50 <.0001
  int_st                     2.2977   1.2180   0.0992   150.64 <.0001
  time                      -0.0730  -0.0457   0.0246     3.45 0.0633
  trt       Placebo          0.1165   0.3037   0.0448    46.04 <.0001
  trt       Progabide        0.0000   0.0000   0.0000      .    .
  time*trt  Placebo          0.0315   0.0066   0.0375     0.03 0.8603
  time*trt  Progabide        0.0000   0.0000   0.0000      .    .
  baseline                   0.0225   0.0378   0.0015   619.46 <.0001
  age                        0.0163   0.0085   0.0027     9.72 0.0018

                  Marginal Mean Parameter Estimates
                     and Sandwich Standard Errors

  Parameter Level1    Initial value Estimate   StdErr    ChiSq Pvalue

  int_s                      0.1055  -1.3907   1.3798     1.02 0.3135
  int_t                     -0.5665  -1.6024   1.4621     1.20 0.2731
  int_st                     2.2977   1.2180   0.5099     5.71 0.0169
  time                      -0.0730  -0.0457   0.0468     0.95 0.3289
  trt       Placebo          0.1165   0.3037   0.2456     1.53 0.2163
  trt       Progabide        0.0000   0.0000   0.0000      .    .
  time*trt  Placebo          0.0315   0.0066   0.1550     0.00 0.9660
  time*trt  Progabide        0.0000   0.0000   0.0000      .    .
  baseline                   0.0225   0.0378   0.0091    17.27 <.0001
  age                        0.0163   0.0085   0.0129     0.44 0.5094
```

Figure 8.2: *Epilepsy Data: Results as output by macro* **GEE2Counts** *after fitting (5.10).*

Chapter 9

The Combined Model as a Correlated or Overdispersed Count Data Simulator for Marginal Models; A SAS Implementation

9.1 Introduction

Huge amounts of data are the result of the vast amount of research going on in so many fields of study. These data need to be analyzed and summarized into meaningful and informative statements.

On the one hand, interest may be in the analysis of these data, which is commonly done using statistical methods that depend on the kind of data at hand. In medical research, for example, it is often the case that each patient has data recorded repeatedly or observed multiply over time. This introduces the phenomenon of correlated data because observations from one patient will be more related than observations across different patients. Molenberghs and Verbeke (2005) and Verbeke and Molenberghs (2000) describe to methods for the analysis of discrete and continuous longitudinal data, respectively.

On the other hand, interest may be in evaluating the statistical prop-

erties underlying certain data generating mechanisms. In this light, Monte-Carlo (MC) simulations are carried out, in which samples are randomly drawn from probability distributions to mimic statistical processes that can be used to study properties of statistical methods. Simulation of correlated Poisson random variables is a topic of ongoing research and various methods have been proposed in the literature to this end, for example, the overlapping sums (implemented in Matlab by Madsen and Dalthorp, 2007; Mardia, 1970; Kocherlakota and Kocherlakota, 1992, 2001); Lognormal-Poisson hierachy, implemented Matlab; normal to anything (NorTA, Cario and Nelson, 1997, 1998; Nelsen, 2006; Mardia, 1970; Li and Hammond, 1975, available in R), and extensions thereof (Yahav and Shmueli, 2012; Ghosh and Pasupathy, 2012; Shin and Pasupathy, 2010; Avramidis *et al.*, 2009; Park and Shin, 1998; Downer and Moser, 2001) mostly implemented in R and Matlab. See also Devroye (1986) for an overview on random variate generation. These tools yield correlated Poisson random variables with the specification of the Poisson means and the desired or target correlation structure. Most of these methods, however, suffer from such limitations as: severe computational restrictions; difficulty achieving the target correlation; generated variables are required to be overdispersed; low correlations obtained; correlations constrained to be strictly positive; etc. Alternatively, random effects can be used to induce correlation, thereby generating data from a hierarchical model. If the simulation is in the context of hierarchical models, this approach would be fine. However, whenever interest is in population-averaged or marginal models, the parameters used in the hierarchical model do not have a simple correspondence with those in the marginal model. Given such a tool as the combined model that incorporates the two common features of count data, namely, overdispersion and correlation, it certainly is essential to generate data from such a method whenever interest is in simultaneously investigating these features.

In the context of continuous or normal longitudinal data, calculations are computationally easier than in the non-normal case because the model for the response variable given random effects is the normal distribution and that of the random effects is the normal distribution as well. The two combined

and integrating over the random effects leads to a normal distribution as the marginal model. In the non-normal case though, the model for the outcome variable and the random effects combined does not lead, in general, to closed-form solutions for the marginal model. Even if it does, expressions tend to be cumbersome. This is due to the lack of elegant and convenient multivariate distributions analogous to the case of longitudinal data that can be assumed normally distributed. This poses computational and interpretational challenges. Specific to count data, which is of interest here, evaluation of the multivariate Poisson distribution grows in computational complexity with an increase in the dimensions due to the summations inherent in the distribution (Karlis, 2003). It is therefore of interest to find alternative means of analysis of correlated count data. One alternative is the generalized linear mixed model (GLMM) proposed by Breslow and Clayton (1993). This model accounts for the correlation by use of effects specific to a subject or study unit (random effects) and then derives the marginal distribution as a result of combining a random-effects distribution with a Poisson distribution for the data given the random effects. Molenberghs *et al.* (2007) and Molenberghs *et al.* (2010) have introduced the so-called combined model (CM) as a tool to model data that is not only correlated but also overdispersed. Overdispersion may occur when the model restricts the data in the sense that the variance expected from the model is less than that observed in the data. It is commonly encountered in data assumed to follow a binomial distribution, correlated or uncorrelated, correlated Bernoulli/binary random variables, correlated or independent observations arising from counting processes (Poisson data) and time-to-event/survival data. This is due to the mean-variance relationship inherent in the distributions that are assumed to be the data generating mechanisms. Overdispersion is, however, not an issue in the case of independent Bernoulli observations. Research has shown overdispersion to be caused by, for example, missing covariates and the presence of correlation between individual responses or clustering, among others. Depending on outcome type and model, not accounting for overdispersion may lead to bias in some or all parameters; it definitely biases precision estimates. The result is then usually

smaller p-values for the statistical tests as well as, of course, confidence intervals that are narrower than should be if overdispersion were properly handled. This means that inference based on such statistical analyses is questionable and may be misleading.

Solutions have been proposed in the literature and implemented in statistical software to account for overdispersion. The negative-binomial (NEGBIN) model for count data is one such tool which assumes the count data to have the Poisson as the parent distribution and a Gamma distribution for the extra parameter that accounts for overdispersion. The resulting marginal distribution is then the negative-binomial distribution. Note that earlier statistical analyses were generally only able to account for either correlation or overdispersion, but not both. But, given data that exhibit both features, it is a necessity to account for both in analyses, indeed. We refer to Section 3 for a detailed description of the GLMM, negative-binomial, and combined models.

We here describe an implementation of the combined model, as discussed in chapter 6, in a SAS 9.3 (SAS Institute Inc., 2011) macro to generate correlated and/or overdispersed Poisson random variables. The method can also be used to generate purely serially correlated counts by dropping the normal random effects and choosing the "overdispersion part" to follow a serially correlated multivariate Gamma distribution. The macro makes use of the SAS **LOGISTIC** procedure to create the design matrix by using a working response variable, which is deleted after it has served its purpose. At a data manipulation phase, the macro also uses the **GENMOD** procedure to obtain parameter estimates corresponding to the design matrix in order to eliminate columns from the design matrix corresponding to reference categories, in case of classification variables and depending on the parameterization method used, for example, glm. By making use of these procedures, any parameterization possible in SAS **LOGISTIC** and **GENMOD** procedures can be used for the design matrix, e.g., effect, glm, ordinal, reference, etc. See SAS documentation about the different parameterization methods for classification variables at `http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect006.htm`. We show, using 2 ex-

amples, namely, a case for the random intercept model and for a random intercept and slope model, how to use the macro to generate data. A review of the models used for correlated and/or overdispersed count data is presented in chapter 3, while chapter 6 details the method on data generation that is implemented in this chapter. Section 9.2 provides some specific details about the SAS macro, some examples of combined models to generate data from, how to generate these data using the macro and the output of the macro. Some concluding remarks are given in Section 9.3.

9.2 The SAS Macro

9.2.1 Introduction

We have implemented the above-discussed method of data generation for marginal models in SAS version 9.3. The SAS macro is called **CorrPoisson**. Data generation is done in SAS/IML preceded by some data manipulations. Given that we allow the θ_{ij}'s to be correlated, some form of multivariate Gamma distribution is required. We invoke the copula package in the freely available R statistical programming language (R Development Core Team, 2012; Hofert *et al.*, 2012; Hofert and Maechler, 2011; Yan, 2007; Kojadinovic and Yan, 2010) to do this. We are aware of the experimental COPULA procedure in SAS9.3 but follow a different route here. Instead, we have explored SAS9.3's flexibility to call R in the SAS/IML procedure using the SUBMIT and ENDSUBMIT statements. This is done in the **RinSASIML.sas** program which is included in **CorrPoisson.sas** using the *%inc* statement. Before running **CorrPoisson**, the path to **RinSASIML.sas** must be defined correctly using the *%let* macro statement as, for example, *%let include=c:/temp;* in case **RinSASIML.sas** is saved in the *c:/temp* directory. Three issues deserve mention at this point, namely, (a) that R software (available at `http://cran.r-project.org.`) has to be installed, (b) that the SAS system has to be launched with the -RLANG system option to permit calling R from SAS *"(it is often convenient to insert this option in a SASV9.CFG file)"*, and, (c) that calling R functions in SAS using the SUBMIT and ENDSUBMIT

Table 9.1: *SAS compatibility with R releases as obtained from Wicklin (2013)'s blog.*

SAS Version	**PROC IML**	SAS/IML Studio	Release Date	R Versions
9.2	N/A	3.2	Jul 2009	2.6.1 - 2.11.1
9.22	9.22	3.3	Nov 2010	2.9.1 - 2.11.1
9.3	9.3	3.4	Jul 2011	2.9.1 - 2.15.3
9.3m2	12.1	12.1	Aug 2012	2.9.1 - 2.15.3
9.4	12.3	12.3	Jul 2013	2.13.0 - 3.0.1
9.4m1	13.1	13.1	Dec 2013	2.13.0 - present

statements is a relatively new feature that was introduced in SAS/IML 9.22. As such, the macro will not work with SAS versions that preceded 9.22. We refer to the SAS/IML® 9.22 User's Guide (SAS Institute Inc., 2011) or later versions for details about calling R from within SAS. The macro was developed and certainly works with SAS version 9.3 and R version 2.14.2. However, compatibility issues may arise while invoking R in SAS/IML depending on the versions of both R and SAS. Error messages that may indicate incompatibility are, for example, "An installed version of R could not be found" or "The installed version of R cannot be used". Table 9.1, obtained from Wicklin (2013)'s blog, presents an overview of the match between the latest SAS versions and the corresponding R releases they support. In general, specific SAS versions support specific sets of R releases. It is necessary that the user first ensures that calling R from SAS is permitted. A quick test is to run the following code in SAS:

```
proc options option=RLANG;
proc iml;
submit /R;
getwd()
endsubmit;
run;
```

If this test results in errors and the user has the most recent version of R, it may

be useful to explicitly tell SAS/IML which R version to use since SAS/IML tries to use the corresponding R_HOME variable. This can be done by launching SAS with a specification of the R_HOME variable as, for example, -RLANG -SET R_HOME "C:/program files/R/R-3.0.1", in which case R version 3.0.1 would be used. We emphasize that in order to have a successful execution of macro **CorrPoisson**, one should first ensure that (1) the connection between SAS and R is ok (by running the test program above), and (2) that the path to **RinSASIML.sas** is correctly defined. Otherwise, errors directly related to these two aspects may be encountered. Table 9.2 presents the arguments of that aid the application of the macro. In what follows, in Sections 9.2.2 and 9.2.3, we illustrate the use of the macro with two cases, namely, (1) when a random-intercept model is specified for the normal random effects and (2) when a random intercept and slope model is used for the normal random effects.

Table 9.2: *The macro arguments for* **CorrPoisson** *and their corresponding description.*

Macro argument	**Description**
`CovData`	Dataset containing subject or cluster identification variable and covariates from which design matrices X_i (and $\widetilde{X}_i$) in (6.1a) is (are) created. By default, `CovData = temp`, for illustration purposes. Dataset `temp` contains covariates `id, trt` and `time`. It is a required argument and therefore has to be specified before running the macro. This dataset should take the "long" or hierarchical data structure as opposed to the wide format. Please note that all variables to be specified in the other macro arguments must be in this dataset.
`ID`	Subject identification variable in `CovData`
	Continued on next page

Table 9.2 – continued from previous page

Macro argument	Description
`OrderVar`	Variable with ordering of the observations within subject. This should be in `CovData` dataset. Default input is `time`, again for illustrating how the macro can be invoked.
`Xcov`	Covariates from which X_i ($\widetilde{X}_i$) is created. It is also a required argument though for illustration purposes, `Xcov = trt time trt*time`. Please note that the intercept must not be included in the specification of `Xcov`. It is added at the creation of the X_i design matrix.
`Alpha`	The desired marginal mean parameters, without the reference categories. It can be specified, for example, as `Alpha = 2.5 0.7 1.2 -0.45`, corresponding to "Intercept", "trt=0", "time" and "trt*time", respectively. Please note that the value for the intercept must always be included.
`Class`	Specify all classification variables included in the `Xcov` argument, for example, `Class = trt`.
`outdata`	Name of final output dataset in which the generated outcome variable, named Y, is merged with the `CovData` dataset. It also contains the variance-covariance matrix of the Gamma distribution (Σ_i, "GammaCov") in (3.16c), the corresponding correlation matrix of the Gamma distribution ("GammaCorr"), the "shape" and "scale" parameters for the Gamma distribution, the θ_i's ("GamV") in (3.16c), the λ_{ij}^*'s ("mu") in 3.16b and the random effects estimates ($\boldsymbol{b}_i$) in (3.16d). By default, `outdata = out`.

Continued on next page

Table 9.2 – continued from previous page

Macro argument	Description
`param`	Parameterization method for the classification variables specified in the `Class` argument. Default is `param = glm`. See SAS documentation for details about the different methods at `http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect006.htm`.
`random`	Covariates for the normal random effects. Default is blank hence random intercept model. To fit, for example, a random intercept and slope model, specify `random = time`, whereby `time` indicates the ordering of observations within a subject.
`meanNormalRE`	By default (`meanNormalRE` is left blank), macro **CorrPoisson** then assumes the normal random effects ($\boldsymbol{b}_i$) to have zero mean. This can be changed by filling the mean in this argument.
`seed`	Set seed. Default is `seed = 123`.
`desiredVarCov`	Specify the desired variance-covariance matrix V by inputting the entries of the upper triangular matrix rowwise (e.g., v11 v12 v22 or v11 v12 v13 v22 v23 v33).
`GammaRandEff`	Either `0`, `1` or `2` for the Gamma random effects where `0`=No Gamma random effects, `1`=Independent Gamma random effects, `2`=Correlated Gamma random effects. Default is `GammaRandEff = 2`.
`NormalRandEff`	Either `0`, `1` or `2` for the normal random effects where `0`=No normal random effects, `1`=Independent normal random effects, `2`=Correlated normal random effects. Default is `NormalRandEff = 2`.

Continued on next page

Table 9.2 – continued from previous page

Macro argument	Description
`Estimates`	Print results to output window, Yes (`1`) or No (`0`). Default is `Estimates = 1`.

9.2.2 The Random Intercept Case

In this section, we demonstrate how to generate correlated count data from the combined model using the macro that we have developed.

9.2.2.1 The Combined Model

Using a random-intercept model for the normal random effects, we illustrate the generation of 4 correlated Poisson variables using the following combined model:

$$\begin{aligned} Y_{ij} &\sim \text{Poi}(\lambda^*_{ij}), \\ \lambda^*_{ij} = \theta_{ij}\lambda_{ij} &= \theta_{ij}\exp(\xi_0 + b_{0i} + \xi_1 T_i + \xi_2 t_{ij} + \xi_3 T_i * t_{ij}), \\ \boldsymbol{\theta}_i &\sim \text{MGamma}(\text{mean} = \mathbf{1}, \text{variance} = \Sigma_i), \\ \boldsymbol{b}_i = b_{0i} &\sim N(0, d), \end{aligned} \tag{9.1}$$

where the treatment allocation $T_i \sim$ Bernoulli(0.5), t_{ij} is the ordering of the j^{th} observation in subject $i = 1, \ldots, K = 500$ and $j = 1, 2, 3, 4$. The design matrices X_i and $\widetilde{X}_i$ in Equation 6.1a are created from the same covariates, namely, T_i, t_{ij}, and $T_i t_{ij}$ while the desired marginal mean parameters and desired variance-covariance structure are

$$\boldsymbol{\alpha} = \begin{pmatrix} 1.521 \\ 0.437 \\ -0.254 \\ 0.145 \end{pmatrix}$$

and

$$V = \begin{pmatrix} 256 & 128 & 144 & 224 \\ 128 & 208 & 228 & 172 \\ 144 & 228 & 299 & 296 \\ 224 & 172 & 296 & 567 \end{pmatrix},$$

respectively.

9.2.2.2 Calling Macro CorrPoisson

The Combined model has several variations, as shown in Table 9.3 and described in Section 6.2.2. Generating data from these variations can be achieved by specifying the combinations of normal and Gamma random effects using the macro arguments `NormalRandEff` and `GammaRandEff`, respectively. For illustration purposes, we shall only generate from the general case of the combined model, namely, with correlated normal and correlated Gamma random effects. This is done by specifying `NormalRandEff = 2` and `GammaRandEff = 2`. See Table 9.3 for the other possibilities. Note that in the case of a random intercept model, setting `NormalRandEff` either equal to `1` or `2` is equivalent because one is dealing with only a single value d. To generate data from (9.1), one has to first create the `CovData` dataset, referred to as `temp` in our example. This dataset contains variables `id, trt` and `time`. See Table 9.2 for more details about the `CovData` argument. Given the `temp` dataset, the following macro call would generate correlated count data from the combined model with a random intercept model specified for the Normal random effects:

```
%CorrPoisson(CovData=temp, id=id, OrderVar=time,
Xcov=trt time trt*time, Alpha=1.521 0.437 -0.254 0.145,
Class=trt, outData=out, random=, GammaRandEff=2,
NormalRandEff=2,
desiredVarCov=256 128 144 224 208 228 172 299 296 567);
```

Generating from the other variations is done by specifying the macro arguments as shown in Tables 9.3 and 9.2. Please note that setting `NormalRandEff`

`= 0` meaning no normal random effects, and `GammaRandEff = 0` meaning no Gamma random effects, would imply generating independent count data which is not of interest in this thesis although it is also implemented in macro **CorrPoisson**, for completeness. Also note that not all macro arguments are shown in the above call. Those not shown are set to the default values. See Table 9.2 for the full list of macro arguments that facilitate the data generation and their descriptions. Leaving the `random` argument blank implies the use of a random intercept model for the normal random effects.

9.2.2.3 Output

By default, **CorrPoisson** creates the output dataset `out` whose content is as described in Table 9.2 under the `outData` argument. Since `Estimates = 1` by default, it also prints to the output window as shown in Figure 9.1. As expected for a combined model with a random intercept only model, a change (diff) in the marginal parameters ($\boldsymbol{\alpha}$) from the hierarchical parameters ($\boldsymbol{\xi}$) is only in the intercept parameter but the other parameters remain practically unchanged.

9.2.3 The Random Intercept and Slope Case

Consider the following combined model with a random intercept and slope model specified for the normal random effects:

$$\begin{aligned} Y_{ij} &\sim \text{Poi}(\lambda^*_{ij}), \\ \lambda^*_{ij} = \theta_{ij}\lambda_{ij} &= \theta_{ij}\exp((\xi_0 + b_{0i}) + \xi_1 T_i + (\xi_2 + b_{1i})t_{ij} + \xi_3 T_i * t_{ij}), \\ \boldsymbol{\theta}_i &\sim \text{MGamma}(\text{mean} = \mathbf{1}, \text{variance} = \Sigma_i), \\ \boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}\right], \end{aligned} \tag{9.2}$$

where notation and the desired marginals are as in Section 9.2.2.1. Since model (9.2) is similar to (9.1) with the only difference being the random slope (b_{1i}) in (9.2), the code shown in Section 9.2.2.2 is used to generate correlated count data. The only difference is that, in this case, the argument `random = time`.

Table 9.3: *Possible combinations of the Normal and Gamma random effects in the context of count data.* ✓ *refers to combinations of the combined model from which correlated and/or overdispersed data can be generated, while* ✗ *refers to the independent count data generation case which is not of interest in this thesis. Also included is the SAS macro argument referring to the Normal and Gamma random effects and the corresponding value to be input for each case, when running macro* **CorrPoisson**.

			Gamma random effects		
			Yes		No
			Correlated	Independent	
Normal random effects		SAS macro argument and corresponding input	`GammaRandEff = 2`	`GammaRandEff = 1`	`GammaRandEff = 0`
Yes	Correlated	`NormalRandEff = 2`	✓	✓	✓
	Independent	`NormalRandEff = 1`	✓	✓	✓
No		`NormalRandEff = 0`	✓	✓	✗

```
        Generate Correlated Poisson data from CM with
          Normal and Correlated Gamma random effects
     ****************************************************

                    Sample size (K) = 500
       minimum number of measurements per subject = 4
       maximum number of measurements per subject = 4
        Normal random effects covariates = Intercept


     Given Mean parameters are: Intercept =     1.521
                                trt0            0.437
                                time           -0.254
                                trt0time        0.145

                                          Y1        Y2        Y3        Y4

Given variance-covariance matrix =  Y1       256       128       144       224
                                    Y2       128       208       228       172
                                    Y3       144       228       299       296
                                    Y4       224       172       296       567


     Parameter      alpha       beta       diff          D

     Intercept      1.521   1.518115   0.002885    0.00577
     trt0           0.437      0.437 -3.13E-14
     time          -0.254     -0.254  1.61E-15
     trt0time       0.145      0.145 -3.47E-15
```

Figure 9.1: *Results printed by macro* **CorrPoisson** *to the output window when a random intercept model is used for the normal effects. alpha and beta in the output are actually* $\boldsymbol{\omega}$ *and* $\boldsymbol{\xi}$*, respectively, following notation from Chapter 6.*

The macro will then create a dataset containing the generated data named `out`, in long form, and a print to the output window as in Figure 9.2. Unlike the random intercept case, the difference between $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ in the case of the random intercept and slope model is evident in both the intercept and time parameters, as expected. However, this pattern holds when all subjects have an equal number of measurements in the `CovData` dataset, in the case that the random intercept and slope model for the normal random effects is specified. Note that for the random intercept model, a change is seen only in the intercept parameter, whether or not all subjects are completers. Not only does the macro allow for n_i to be equal but also allows n_i to be unequal. To illustrate this, we generate a dataset with n_i being between 2 and 4 measurements

```
        Generate Correlated Poisson data from CM with
    Correlated Normal and Correlated Gamma random effects
     ****************************************************

                    Sample size (K) = 500
       minimum number of measurements per subject = 4
       maximum number of measurements per subject = 4
     Normal random effects covariates = Intercept time


      Given Mean parameters are: Intercept =     1.521
                                 trt0            0.437
                                 time           -0.254
                                 trt0time        0.145

                                          Y1        Y2        Y3        Y4

Given variance-covariance matrix =  Y1       256       128       144       224
                                    Y2       128       208       228       172
                                    Y3       144       228       299       296
                                    Y4       224       172       296       567


  Parameter     alpha      beta      diff         D

  Intercept     1.521 1.5195865 0.0014135 0.0040014 0.0000601
  trt0          0.437     0.437 -1.59E-14 0.0000601 0.0002349
  time         -0.254 -0.254647 0.0006473
  trt0time      0.145     0.145 -1.86E-15
```

Figure 9.2: *Results printed by macro* **CorrPoisson** *to the output window when a random intercept and slope model is used for the normal effects when the subjects have equal number of measurements. alpha and beta in the output are actually* $\boldsymbol{\omega}$ *and* $\boldsymbol{\xi}$*, respectively, following notation from Chapter 6.*

per subject. For one to fit the same model shown in Section 9.2.3 but with n_i varying between 2 and 4 measurements, the same code as in the above call would be used. The difference, though, should be in the input dataset `temp` specified using the `CovData` argument. The results as printed to the output window as shown in Figure 9.3. One should note the difference in the two models from the minimum and maximum number of measurements per subject.

```
        Generate Correlated Poisson data from CM with
     Correlated Normal and Correlated Gamma random effects
      ****************************************************

                    Sample size (K) = 500
          minimum number of measurements per subject = 2
          maximum number of measurements per subject = 4
       Normal random effects covariates = Intercept time


        Given Mean parameters are: Intercept =     1.521
                                   trt0            0.437
                                   time           -0.254
                                   trt0time        0.145


                                            Y1        Y2        Y3        Y4

Given variance-covariance matrix =  Y1      256       128       144       224
                                    Y2      128       208       228       172
                                    Y3      144       228       299       296
                                    Y4      224       172       296       567


   Parameter     alpha      beta      diff         D

   Intercept     1.521 1.5209283 0.0000717 0.0071187 -0.001993
   trt0          0.437 0.4370443 -0.000044 -0.001993 0.0016015
   time         -0.254 -0.255718 0.0017181
   trt0time      0.145 0.1449747 0.0000253
```

Figure 9.3: *Results printed by macro* **CorrPoisson** *to the output window when a random intercept and slope model is used for the normal effects and there are varying number of measurements per subject. alpha and beta in the output are actually* $\boldsymbol{\omega}$ *and* $\boldsymbol{\xi}$*, respectively, following notation from Chapter 6.*

9.3 Concluding Remarks

We have presented SAS code to generate correlated Poisson data, in the context of marginal models, from the combined model introduced by Molenberghs *et al.* (2007, 2010). The combined model simultaneously accommodates correlation and overdispersion unexplained by the normal random effects. In the absence of correlation, the model simplifies to a negative-binomial model for overdispersion. On the other hand, in the absence of overdispersion, it simplifies to the GLMM. The model's flexible structure makes it a good candidate as a data generator reflecting the characteristics of interest, in this case,

overdispersion and/or correlation. The CM is a convenient tool that mimics or incorporates these intrinsic features of correlated count data.

By marginalizing the distribution of the Poisson response conditional on the normal and Gamma random effects i.e., integrating out the random effects from the conditional density of the combined model, one is able to generate data in the context of marginal models by comparing the mean and variance of the marginal model with desired marginal mean and variance structures. The covariates determining the fixed- and random-effects design matrices are kept simple herein. This is not limiting in the sense that a specification of any covariates can be done as is needed. It is possible to encounter non-positive definite D matrices or negative entries along the diagonal of Σ_i. This may point to a non-allowable hierarchical model to come with the marginal model or perhaps a marginal model that is in itself not allowable. The analogy would be a multivariate normal with a given but non-positive definite variance-covariance matrix. Such model is invalid in the first place and needs to be reconsidered.

Because the combined model is hierarchical, random variables with only positive correlations are generated due to restrictions of positive-definiteness on the random effects variance-covariance matrices. This may be a drawback for the combined model, as is the case for some of the methods present in the literature for count data generation. However, a way to overcome this is to generate directly from the marginal model, arguably via correlated θ_{ij}, of which the variance-covariance matrix Σ_i then reflects the desired structure.

Execution errors of the form "Unable to allocate sufficient memory" may be encountered when attempts are made to generate sizes of datasets that use resources more than are available to SAS for the specific computer in use. One may then have to reduce the sample size or the number of measurements per subject or find a computer with greater memory capacity. Our experience is that this limitation is more rampant in 32-bit Windows operating systems which support matrices up to a maximum size of memory of 2GB of addressable space.

The SAS macro **CorrPoisson.sas** and the **RinSASIML.sas** program

are available at the authors' website (`http://ibiostat.be/software/overdispersion`).

Chapter 10

Estimating the Random Effects Distribution of Linear Mixed Models using SAS

10.1 Introduction

The linear mixed (effects) model is a generalization of the standard linear model where data are allowed to exhibit correlation and nonconstant variability. It is the routine framework of analysis for longitudinal data (subject responses are repeatedly measured over time) but can very easily be adopted for other data structures like clustered data (correlated within cluster), multivariate data (several responses measured for each experimental unit or subject), etc.

For the linear mixed model, the mean of the response is linear in terms of certain parameters and incorporates both fixed and random effects. The fixed effects are the population-averaged parameters associated with known explanatory variables while the random effects are subject-specific parameters associated with randomly drawn subjects from a population. One common phenomenon with longitudinal data is that different subjects exhibit different patterns of evolution. Some subjects may start evolving below or above the

average starting point (population-averaged intercept) with a rate of evolution faster or slower than the average rate (population-averaged slope). This phenomenon is accounted for in the linear mixed model by the random effects. They reflect the between-subject variation or the deviation of the subject-specific evolution from the average evolution.

The random effects and the measurement error term are usually assumed to be normally distributed. Butler and Louis (1992) and Verbeke and Lesaffre (1997) show that inference on the fixed effects is robust to nonnormality of the random effects. However, the assumption of normality may be too restrictive and may not yield an efficient estimation of the fixed effects and model-based standard errors. A deviation of the random effects distribution from normality may also affect inference on the random effects. Verbeke and Lesaffre (1996) show that the empirical Bayes estimates are forced to satisfy normality even when the underlying distribution is not normal but a mixture of normals. This then necessitates a much more flexible assumption for the random effects distribution when fitting a linear mixed model.

As introduced in Section 3.3.3, Ghidey *et al.* (2004) proposed the penalized Gaussian mixture linear mixed model as an alternative that can be used to estimate a more flexible random effects distribution. This method fits a linear mixed model but with a more flexible and general distribution function for the random effects. They implemented this method in MATLAB but their software is not publicly available, to-date, for use should one be interested in fitting the PGM model. We hereby implement it in a SAS macro which makes it easy and user friendly for SAS users to fit this model. The SAS macro takes the initial values directly from the SAS/STAT procedure `MIXED` and also creates the design matrix for the fixed effects using the SAS/STAT procedure `GLMMOD` which creates dummy variables in case of categorical covariates. `GLMMOD` also deletes observations in case of missing values in the design matrix. Overviews of the classical and the PGM (Ghidey *et al.*, 2004) linear mixed models have been given in Sections 3.3.2 and 3.3.3, respectively. Section 10.2 describes the SAS macro and illustrates, using simulated data, how to fit a random intercept model and a random intercept and slope model. In section 10.3, we analyze

data from the Jimma infant study previously analyzed by Lesaffre *et al.* (1999) and briefly introduced in Section 2.6. Some conclusions and a discussion of our experience with the PGM macro are provided in Section 10.4.

10.2 The SAS Macro

We have implemented the penalized Gaussian mixture linear mixed model in a SAS macro which we have called PGM. The macro uses a two-step iterative procedure, as proposed by Ghidey *et al.* (2004), to estimate the unknown parameters, as described in Section 3.3.3. In step 1, $\mathbf{a}$ is estimated by maximizing (3.11) while conditioning on initial values for the other parameters in $\boldsymbol{\theta}$ obtained from fitting the classical linear mixed model (3.5) using the SAS/STAT procedure MIXED. The second step then updates the other parameters in $\boldsymbol{\theta}$ conditioning on the updated $\mathbf{a}$ from step 1. The $\boldsymbol{\sigma_R}$ vector in $\boldsymbol{\theta}$ is kept fixed (to the estimates from the MIXED procedure) during maximization and is only updated afterwards. A Newton Raphson optimization algorithm using the SAS/IML function NLPNRR is employed at each conditional step. Iteration is done between steps 1 and 2 until convergence.

The unknown parameters $\boldsymbol{\theta}$ in (3.11) are estimated in SAS/IML. It was developed and tested in SAS version 9.1.3 (SAS Institute Inc., 2002-2004) but should work with later versions without any problem. PGM can be used to fit either a random intercept model or a random intercept and slope model. Table 10.1 presents the macro arguments that are required for PGM to execute. To fit a random intercept and slope model, the argument RANDOM should be set to the variable indicating the ordering of measurements within a subject. If left blank (the default), PGM fits a random intercept model. PGM uses unstructured (UN) as the default covariance structure for the random effects, to be used by the procedure MIXED. Specifying another structure can be done (as required by the MIXED procedure) using the argument TYPERAND. The arguments LAMBDA1 and LAMBDA2 specify the vector of smoothing parameters in dimension 1 ($\boldsymbol{\lambda}_1$) and dimension 2 ($\boldsymbol{\lambda}_2$), respectively. By default, `LAMBDA1 = LAMBDA2 = (0.01, 0.1, 1,10,100,1000)`. When fitting a random intercept

model, LAMBDA2 is set to zero. The lower and upper bounds of the interval over which the grid points are spread is specified by the arguments DMIN and DMAX, respectively. By default, DMIN = -4 and DMAX=4. The maximal number of iterations for convergence is specified using the argument MAXITER(= 20,000 by default). J and L are arguments that specify the number of grid points in dimensions 1 and 2, respectively. When fitting a random intercept model, L is set equal to 1. By default, J = L = 20 for a random intercept and slope model. Ghidey *et al.* (2004) recommend using a penalty order difference e in (3.11) equal to 3, which is the default in macro PGM.

Table 10.1: *Required macro arguments in order to run macro PGM.*

DATA	=	Input data set structured as required by the SAS/STAT procedure MIXED.
SUBJECT	=	The variable with identification numbers of the subjects in the data set.
CLASS	=	Class variables in the model excluding SUBJECT. PGM considers SUBJECT as a class variable by default.
RESPONSE	=	The response variable.
FIXED	=	The fixed effects (variables) in the model. Note that the intercept is included by default. For example, if your model is: $\boldsymbol{Y}_i = \xi_0 + \xi_1 * time + \xi_2 * gender + \xi_3 * (gender * time) + \boldsymbol{\epsilon}_i$, then, `FIXED = time gender gender*time.`

To specify either a 1 or 2 as the penalty order difference, arguments PORD1 and PORD2 can be used for dimension 1 and 2, respectively. When Empirical Bayes (EB) estimates are of interest, they are saved in a data set named by the argument OUTEB. By default, no name is given hence the EB estimates are not calculated. The λ(s) that minimize(s) AIC (MINLAM1 and MINLAM2), the corresponding degrees of freedom (MINDF), minimum AIC (MINAIC), penalized log-likelihood (MINLLP) and **a**-coefficients are stored by default, in a data set called "OPTIMAL" using the argument OUTOPT. Argument OUTSOLF is used

to name the fixed effects solution data sets. By default, `OUTSOLF = SOLUTIONF` leads to the data sets "SOLUTIONF_MIXED" from the classical linear mixed model and "SOLUTIONF_PGM" for the PGM model. Argument `OUTCOV` is used to name the data set for the covariance parameters (elements of the D matrix for the random effects and the residual variability (σ^2). By default, `OUTCOV = CovParms`. `PGM` then outputs 2 data sets called "COVPARMS_MIXED" for the classical linear mixed model and "COVPARMS_PGM" for the PGM linear mixed model. The plots of the estimated random effects distribution are output by default using the argument `PLOTS`. If left blank, no plots will be created. Other arguments are: `ABSGTOL` which specifies the absolute gradient stopping criterion (by default, equal to 10^{-5}), `FUNC_CALL` which specifies the maximum number of function calls in the Newton Raphson optimization process (default is $10,000$) and `DENSITY` to either retain the data sets containing the estimated random effects density values used for plotting or not. When fitting a random intercept and slope model, these data sets are: "DENSITY" containing the joint density values, "INTERCEPT" containing the marginal density values of the random intercept and "SLOPE" containing the marginal density values of the random slope. When fitting a random intercept model, the data set is called "INTERCEPT". By default, the density data sets are deleted (argument **DENSITY** = NO). To retain them, set `DENSITY = YES`. Intermediate data sets including: _SolutionF1, _freq, _covparms1, _coveffectsnames, _Rr, _solfixed, _designmatrix, _labels, _Parm, _data, _design and _optim are created during `PGM`'s execution and deleted before the end of execution. All the output data sets can be found in the work library in SAS.

10.2.1 A Case for the Random Intercept Model

We generated data for 500 subjects assuming the following random intercept model;

$$\mathbf{Y}_{ij} = \xi_0 + b_{0i} + \xi_1 * age_i + \xi_2 * sex_i + \xi_3 * time_{ij} + \varepsilon_{ij} \tag{10.1}$$

where $i = 1, \ldots, 500$, $\jmath = 1, \ldots, n_i \leq 5$, $age_i \sim N(60, 1)$, sex_i(male or female)$\sim bin(0.5)$, $time_{ij}$ is the ordering of the $\jmath$-th observation within subject i (each subject was with equal probability allowed to have 1 to 5 measurements), $\mathbf{Y}_{ij}$ is the response of subject i at time point $\jmath$, the random intercept $b_{0i} \sim 0.5 \times N(-1, 0.25^2) + 0.5 \times N(1, 0.25^2)$, the error term $\varepsilon_{ij} \sim N(0, 0.4^2)$ and $\xi_0 = 1, \xi_1 = 0.7, \xi_2 = 0.2, \xi_3 = 1$. We then fitted model (10.1) using the PGM approach with $b_{0i} \sim \sum_{j=1}^{J=20} c_j N(R\mu_j, RD_sR)$. Notice that `RANDOM` is left blank when fitting this model.

10.2.1.1 Sample Call of Macro `PGM`

The macro call to fit a random intercept model is:

```
%PGM(DATA =input1, SUBJECT =id, CLASS= sex, RESPONSE =y,
FIXED =age sex time, RANDOM = );
```

10.2.1.2 Sample Output

The following is output by macro `PGM` when fitting a random intercept model:

1. Printed to the SAS output window is each combination of λ_1 and λ_2 (`LAM1` and `LAM2`), the corresponding degrees of freedom (DF), AIC, log-likelihood (LL), penalty, the penalized log-likelihood (LLP), and whether convergence criteria were met at the two conditional steps (CONVERGED = yes/no). This output is shown in Table 10.2. Note that `LAM2 = 0` and therefore MINLAM2=0 because when fitting a random intercept model, λ_2 in (3.11) equals 0.

2. λ_1, λ_2 that minimize AIC (MINLAM1, MINLAM2, respectively), the corresponding log-likelihood (LOG_LIK), degrees of freedom (MINDF) and minimum AIC (MINAIC) are also printed to the SAS output window as shown below. The method of Gray (1992), for some $\boldsymbol{\lambda}$ however, results in strange df values as can be seen in Table 10.2 affecting the AIC. To circumvent this problem, MINLAM1 and MINLAM2 are chosen to

Table 10.2: *Generated data, random intercept model; Sample output from the PGM model (PEN=PENALTY, CON=CONVERGED).*

LAM1	LAM2	DF	AIC	LL	PEN	LLP	CON
0.01	0	12.13	3036.61	−1506.17	0.3367	−1506.51	Yes
0.1	0	9.97	3034.28	−1507.16	1.1574	−1508.32	Yes
1	0	3.41	3025.75	−1509.46	3.8501	−1513.31	Yes
10	0	−46.37	2944.51	−1518.63	9.2571	−1527.88	Yes
100	0	994.59	5073.59	−1542.21	8.9839	−1551.20	Yes
1000	0	249.45	3621.97	−1561.54	2.6594	−1564.20	Yes

guarantee a) minimum AIC, b) convergence at the two steps during maximization and c) the corresponding df being greater than a *function of the order of the penalty* + the number of free parameters in the model. Constraint c) is motivated by Eilers and Marx (1996) who have shown in the case of uni-dimensional smoothing that the effective dimension (df) as determined by the trace of the smoother matrix approaches the order of the penalty as λ becomes large. Hence, from Table 10.2, the optimal AIC is as shown in Table 10.3.

Table 10.3: *The optimal smoothing parameters combination selected from Table 10.2.*

MINLAM1	MINLAM2	LOG_LIK	MINDF	MINAIC
0.1	0	−1507.16	9.97	3034.28

3. Plot of the estimated random effects distribution as shown in Figure 10.1.

4. Five output data sets including two for the fixed effects solution, two for the covariance parameters and one containing MINLAM1, MINLAM2, MINDF, MINAIC, MINLLP and the estimated **a**-coefficients.

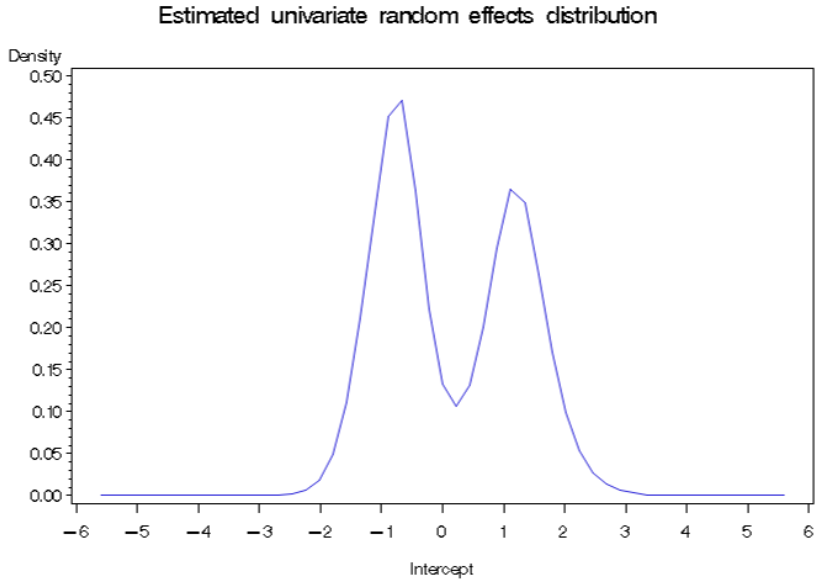


Figure 10.1: *Generated data: Estimated random intercept distribution from PGM linear mixed model (10.1).*

10.2.2 A Case for the Random Intercept and Slope Model

We assumed the following random intercept and slope model and generated data for 100 subjects;

$$\mathbf{Y}_{ij} = \xi_0 + b_{0i} + \xi_1 * age_i + \xi_2 * sex_i + (\xi_3 + b_{1i}) * time_{ij} + \varepsilon_{ij} \qquad (10.2)$$

where $age_i, sex_i, time_{ij}, Y_{ij}, \varepsilon_{ij}$ and $\xi_0, \xi_1, \xi_2, \xi_3$ are as defined in (10.1) and $(b_{0i}, b_{1i})^\top \sim \left[0.5 \times N\left((-1,-0.8)^\top, \mathbf{D}\right) + 0.5 \times N\left((1,0.8)^\top, \mathbf{D}\right)\right]$ are the random intercept and slope with $\mathbf{D} = \begin{pmatrix} 0.25 & 0.1 \\ 0.1 & 0.25 \end{pmatrix}$.

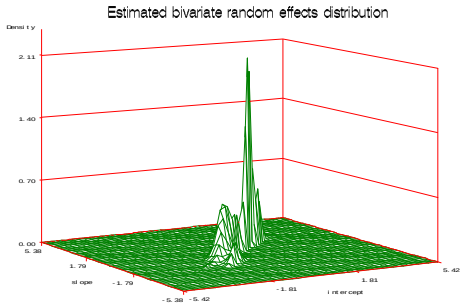
10.2.2.1 Sample Call of Macro `PGM`

A call of the `PGM` macro to fit the PGM linear mixed model for a random intercept and slope model is:

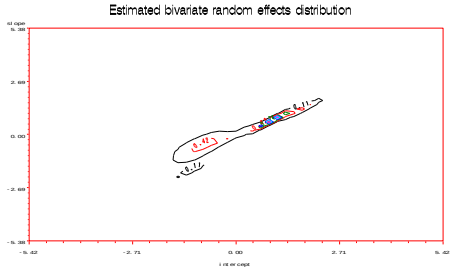
```
%PGM(DATA =input2, SUBJECT =id, CLASS= sex, RESPONSE =y,
FIXED =age sex time, RANDOM =time);
```

10.2.2.2 Sample Output

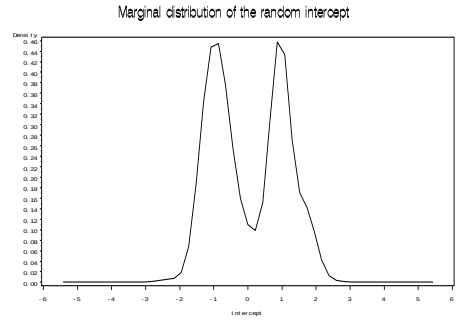
The output of a random intercept and slope model is similar to that described and shown in Section 10.2.1.2 for the random intercept model, with some differences as seen in Table 10.4 and Figure 10.2. The optimal λ_1 and λ_2 and the corresponding log-likelihood, df and AIC are shown Table 10.5.



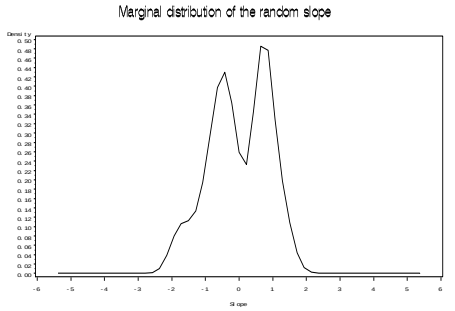
(a) Joint distribution



(b) Contour plot



(c) random intercept



(d) random slope

Figure 10.2: *Generated data: Estimated random intercept and slope distribution from PGM linear mixed model (10.2).*

Table 10.4: *Generated data, random intercept and slope model: Sample output from the PGM model (PEN=PENALTY, CON=CONVERGED).*

LAM1	LAM2	DF	AIC	LL	PEN	LLP	CON
0.01	0.01	20.91	739.81	−348.99	1.7246	−350.72	Yes
0.01	0.1	15.67	730.96	−349.81	1.7601	−351.57	Yes
0.01	1	14.06	728.99	−350.43	1.4401	−351.87	Yes
0.01	10	13.43	728.62	−350.88	1.6957	−352.57	Yes
0.01	100	12.54	727.58	−351.25	1.7118	−352.96	Yes
0.01	1000	13.43	729.68	−351.41	1.6833	−353.10	Yes
0.1	0.01	13.82	729.23	−350.79	1.2067	−352.00	Yes
0.1	0.1	10.50	724.89	−351.95	1.9180	−353.86	Yes
0.1	1	12.30	720.82	−348.11	5.0801	−353.19	Yes
0.1	10	10.67	713.79	−346.22	7.7336	−353.95	Yes
0.1	100	12.30	732.10	−353.75	0.8932	−354.65	Yes
0.1	1000	14.49	736.68	−353.85	0.8262	−354.67	Yes
1	0.01	11.83	726.58	−351.46	0.6482	−352.11	Yes
1	0.1	11.34	727.49	−352.41	1.2322	−353.64	Yes
1	1	11.44	729.78	−353.45	1.3460	−354.80	Yes
1	10	10.41	729.94	−354.57	1.4133	−355.98	Yes
1	100	12.81	736.46	−355.42	1.2157	−356.64	Yes
1	1000	12.65	719.77	−347.24	9.6693	−356.91	Yes
10	0.01	11.75	726.53	−351.51	0.6707	−352.18	Yes
10	0.1	12.89	730.35	−352.29	0.1663	−352.45	Yes
10	1	11.58	728.29	−352.56	0.3136	−352.88	Yes
10	10	10.99	728.88	−353.45	0.0611	−353.52	Yes
10	100	10.98	729.04	−353.54	0.0106	−353.56	Yes
10	1000	10.99	729.09	−353.56	0.0034	−353.56	Yes
100	0.01	11.72	726.49	−351.53	0.6758	−352.20	Yes
100	0.1	11.53	728.97	−352.95	0.8910	−353.84	Yes
100	1	11.47	728.13	−352.60	0.3046	−352.90	Yes
100	10	11.02	727.70	−352.83	0.7059	−353.54	Yes
100	100	11.00	729.10	−353.55	0.0065	−353.56	Yes
100	1000	11.00	729.12	−353.56	0.0011	−353.56	Yes
1000	0.01	11.91	728.13	−352.16	0.4511	−352.61	Yes
1000	0.1	11.53	728.99	−352.97	0.8908	−353.86	Yes
1000	1	11.45	728.11	−352.60	0.2981	−352.90	Yes
1000	10	11.02	729.03	−353.49	0.0489	−353.54	Yes
1000	100	11.00	729.12	−353.56	0.0051	−353.56	Yes
1000	1000	11.00	729.12	−353.56	0.0007	−353.56	Yes

Table 10.5: *The optimal smoothing parameters selected from the results in Table 10.4.*

MINLAM1	MINLAM2	LOG_LIK	MINDF	MINAIC
0.1	10	−346.22	10.67	713.79

10.3 Application to the Jimma Study

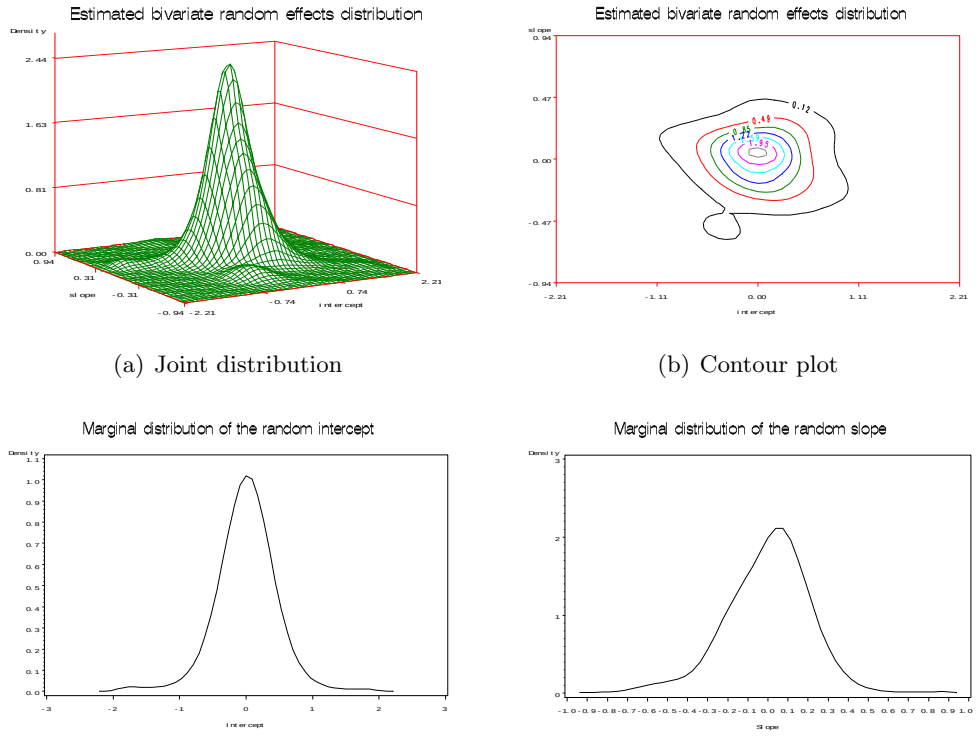
We analyzed data from the Jimma infant study (Lesaffre *et al.*, 1999), introduced in Section 2.6, with interest in estimating the random effects distribution and determining the effect of missing covariates on the estimated random effects distribution. The aim of the Jimma infant study was to identify the determinants of the growth of children in Ethiopia in terms of body weight (kg) as an indicator of health status. The study examined live births from 11 September, 1992 to 10 September, 1993 in one urban area and several rural areas in South-West Ethiopia. It involved 8,000 households and the children were followed up approximately every two months starting immediately after birth over a study period of one year. We considered 495 children ("a 'learning' random sample", Lesaffre *et al.*, 1999) with 3070 observations and fitted the following random intercept and slope model

$$Weight_{ij} = (\xi_0 + b_{0i}) + (\xi_1 + b_{1i}) * agenew_{ij} + \varepsilon_{ij} \tag{10.3}$$

where $i = 1, \ldots, 495$, $j = 1, \ldots, n_i \leq 7$, $Weight$ is the j-th weight of child i observed approximately every two months, $agenew_{ij}$ is the j-th age (in days) of child i transformed by Lesaffre *et al.* (1999) as $agenew_{ij} = \sqrt{a}ge_{ij} - \log(age_{ij} + 1) - 0.02 \times age_{ij}$ motivated by a non-linear evolution of the growth curves of the children, and the distribution of the random effects is given by

$$(b_{0i}, b_{1i})^\top \sim \sum_{j=1}^{J} \sum_{l=1}^{L} c_{jl} N(\boldsymbol{R}\boldsymbol{\mu}_{jl}, \boldsymbol{R}\mathbf{D}_s\boldsymbol{R}^\top). \tag{10.4}$$

The estimated random effects distribution was as shown in Figure 10.3. This



(a) Joint distribution

(b) Contour plot

(c) random intercept

(d) random slope

Figure 10.3: *Jimma infant study: Estimated random effects distribution from PGM linear mixed model (10.3).*

model yielded a minimum AIC of 5343.29 with $\lambda_1 = 1$ and $\lambda_2 = 0.1$ and a corresponding log-likelihood value of -2662.07 and 9.58 df. As can be seen from Figure 10.3, the estimated distribution of the random effects before correcting for baseline covariates is non-normal, especially with regard to the random slope. To correct for baseline covariates, we fit the model

$$Weight_{ij} = (\xi_0 + b_{0i}) + (\xi_1 + b_{1i}) * agenew_{ij} + \boldsymbol{x}_i^\top \boldsymbol{\xi} + \varepsilon_{ij} \tag{10.5}$$

whose random effects distribution is as defined in (10.4). $\boldsymbol{x}_i$ is a vector of child specific baseline covariates which include: age of the mother (*agem*), *sex* of the

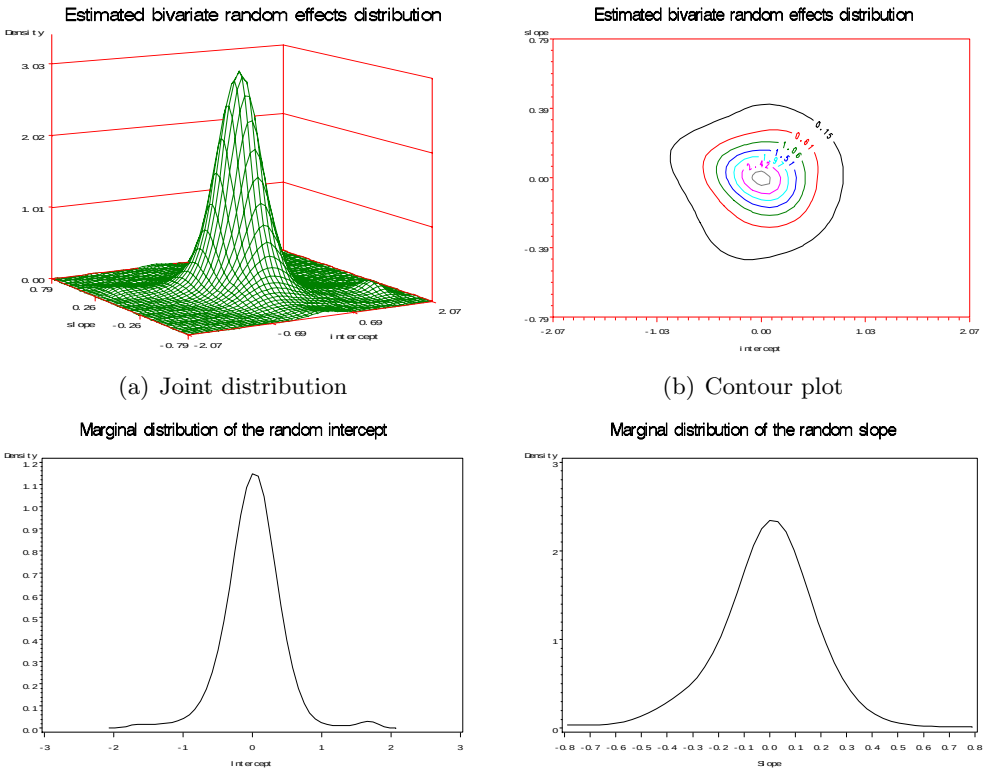
child (male(1) or female(2)), place of delivery of the child (*deliv*=hospital(1), home(2) or health center(3)), level of education of mother (*educ*=illiterate(0), read and write(1), elementary school(2), junior high school(3), high school(4) or college and above(5)), whether mother received antenatal visits during pregnancy(1) or not(0) (*anv*), the period during which child was born (*m1_2*=September-April(0) or May-August(1)) and the interaction of these covariates with the age (*agenew*) of the child; and $\boldsymbol{\xi}$ is a vector of fixed effects parameters corresponding to the baseline covariates. The estimated distribution for the random effects is shown in Figure 10.4. $\lambda_1 = \lambda_2 = 1$ were chosen to be the combination of the smoothing parameters at which AIC was minimal (equal to 5278.05) with a log-likelihood value of -2607.94 and 31.09 df. The fixed effects parameter estimates and their corresponding standard errors for model (10.5) are shown in Table 10.6.

10.4 Concluding Remarks

All models in this chapter were fitted at the default values of the macro arguments. Indeed, the PGM approach quite flexibly estimates the random effects distribution. There is no strict rule as to which number of grid points yield the best estimate of the distribution. In our experience with macro `PGM`, though, a grid (J$\times$ L) of 20×20 would, in many cases, yield quite a good estimate of the random intercept and slope distribution. We however recommend that the distribution be searched over different grid sizes say $10 \times 10, 15 \times 15$ and 20×20 to determine a stable estimate of the distribution. Note though that increasing grid size results in computational intensity depending on the size of the data set, the number of parameters in the model and the number of λ combinations over which to search for the combination that minimizes AIC. When extreme λ values are chosen to minimize AIC, one should re-specify values using the macro argument(s) `LAMBDA1` (and `LAMBDA2`) in the direction of the chosen λ(s) and re-run the program in order to obtain a potential better choice. For example, if, in the case of the random intercept model and basing on the default values for $\boldsymbol{\lambda}_1 = (0.01, 0.1, 1, 10, 100, 1000)$, 0.01 is chosen to min-

Table 10.6: *Jimma infant study: Parameter estimates and standard errors for model (10.5) as output by macro* `PGM`.

EFFNAME	SEX	DELIV	EDUC	ANV_C2	M1_2	Estimate	StdErr
Intercept						2.6042	0.1590
AGENEW						0.9623	0.0517
AGEM						0.0138	0.0037
SEX	1					0.2019	0.0506
SEX	2					0.0000	.
DELIV		1				−0.0654	0.0777
DELIV		2				−0.1525	0.0732
DELIV		3				0.0000	.
EDUC			0			0.0875	0.1451
EDUC			1			0.1164	0.1535
EDUC			2			0.1233	0.1454
EDUC			3			0.0712	0.1468
EDUC			4			0.0242	0.1439
EDUC			5			0.0000	.
ANV_C2				0		−0.0015	0.0557
ANV_C2				1		0.0000	.
M1_2					0	−0.1685	0.0532
M1_2					1	0.0000	.
AGENEW*SEX	1					0.0616	0.0189
AGENEW*SEX	2					0.0000	.
AGENEW*EDUC			0			−0.1611	0.0541
AGENEW*EDUC			1			−0.1425	0.0544
AGENEW*EDUC			2			−0.0997	0.0539
AGENEW*EDUC			3			−0.0668	0.0545
AGENEW*EDUC			4			−0.0439	0.0513
AGENEW*EDUC			5			0.0000	.
AGENEW*DELIV		1				0.0571	0.0321
AGENEW*DELIV		2				0.0050	0.0291
AGENEW*DELIV		3				0.0000	.
AGENEW*ANV_C2				0		−0.0686	0.0207
AGENEW*ANV_C2				1		0.0000	.



(a) Joint distribution

(b) Contour plot

(c) random intercept

(d) random slope

Figure 10.4: *Jimma infant study: Estimated random effects distribution from PGM linear mixed model (10.5).*

imize AIC, then, respecify $\boldsymbol{\lambda}_1 =$ say $(0.0001, 0.001, 0.01)$ (including 0.01 since it may indeed be λ_1 that minimizes AIC) and re-run the macro. Be reminded however, that at the default grid size, computation can get quite intensive and long depending on the size of the data set and the mean structure. For example, the random intercept and slope model for the Jimma infant study took 10 hours without baseline covariates and one week while correcting for baseline covariates. Note that at a specific `LAMBDA1` (and `LAMBDA2`), negative degrees of freedom may result in an error message and a failure of `PGM` to successfully execute. A solution is to specify a grid of values over which to search for `LAMBDA1` (and `LAMBDA2`) that minimize(s) AIC. The method, in general, is computer-intensive. However, we are working on possibilities to improve computation time by, for example, incorporating the computationally intensive part of the method in C++. Macro `PGM` and the updates, when available, can be downloaded from `http://ibiostat.be/software/longitudinal`.

Part IV

General Conclusions, Limitations and Future Research

Chapter 11

General Conclusions and Future Research

11.1 Introduction

This thesis makes a contribution to statistical methodology for the analysis of medical data as well as provide SAS software for the application of the proposed methods. Focus was on important drawbacks present in existing methods, and some solutions were proposed. This chapter provides a general discussion of our findings, lists some limitations in our proposals and suggests areas for future research.

11.2 General Discussion

Methods for the analysis of data stemming from scientific investigations have been presented and described in detail. Applications of the methods in this thesis have been specific to medical studies although the concepts can be applied to other domains of research producing data with similar characteristics. Specifically, Chapter 3 introduced the various existing methods that deal with univariate as well as correlated data elaborating as well on such features as overdispersion. These existing methods suffer from limitations some of which

have been the goal of this thesis to deal with.

In Chapter 4, a proposal is made of a marginal model for longitudinal count data from which inference on both the marginal mean and the association structure is permitted. GEE, one of the traditional approaches in the context of marginal models, permit inference only on the marginal mean but not on the association because it allows that the association structure be misspecified and is therefore a nuisance. To this end, we were not aware of a marginal model, let alone the software implementation of such a method, for longitudinal count data that could be used should scientific interest be not only in the marginal models but also in the association structure. Our proposal in Chapter 4 builds on ideas of pseudo-likelihood to develop a marginal model for correlated count data. A software implementation of this method is presented in Chapter 7. Pseudo-likelihood is known to reduce the computational burden in many applications in which using the joint distribution implies a severe influence on calculations. A comparison of our pseudo-likelihood approach to GEE revealed similarities in the results while pseudo-likelihood offered an additional covariance parameter that quantifies the relationship between each pair of a subject's measurements.

Chapter 5 proposes estimating equations along the lines of GEE in which we build the bivariate Poisson distribution into the score equation at each pair of a subject's measurements. Our proposal permits inference on the marginal mean and covariance parameters, is computationally easier than would be with the full specification of the multivariate Poisson distribution and permits the modeling of the covariance between 2 observations of a subject using covariates.

In order to study of the performance of the proposals in Chapters 4 and 5 via Monte-Carlo simulations, it was necessary to generate correlated count data that reflects the structure of interest, namely, that the parameters generating the date should have a marginal interpretation to facilitate calculation of such quantities as the bias. A common approach is to use a hierarchical model (the generalized linear mixed model) which would induce correlation via the random effects. While this approach would be fine if the context un-

der consideration were hierarchical models, it is certainly not the appropriate way of generating data for marginal model considerations. This is because the parameters from a GLMM do not have a direct marginal interpretation except with extra calculations. Other approaches in literature are either limited in one way or the other, are not implemented in literature or need further work to incorporate covariates especially in terms of the software. The combined model proposed by Molenberghs *et al.* (2007, 2010) turns out useful beyond just modeling correlated and overdispersed data simultaneously but also in the context of data generation. We therefore used this combined model such that with a pre-specification of the desired marginal mean, possibly in terms of covariates, and covariance structure, correlated count data can be generated.

To make application in SAS possible and easy, the 3 proposals in Chapters 4, 5 and 6 were implemented in SAS software in Chapters 7, 8 and 9, respectively. Here, SAS macros were developed that can be very easily run. These macros capitalize on existing SAS procedures like **LOGISTIC** and **GENMOD** to simplify the creation of the design matrix in the presence of classification variables. Thus, our macros can use any parametrization method permitted by these procedures, e.g., effect, glm, ordinal, reference, etc. and the user need not worry about the manual creation of dummy variables or interactions since this process is automated.

Finally, Chapter 10 presented a SAS implementation of the approach of Ghidey *et al.* (2004) to modeling continuous longitudinal data using a linear mixed model that more flexibly estimates the random effects distribution. It is our hope that this software implementation will aid researchers with interest in applying the PGM linear mixed model thereby redeeming time that would be otherwise spent in programming the method were it not available in software.

11.3 Limitations and Further Research

As with many other approaches in literature, there is room for improvement in our proposals and implementations. In Chapter 4, the covariance term was assumed constant across all subjects pairs implying a correlation structure

similar to the exchangeable structure. It is similar due to the dependence of the variance on the mean implying that the correlations obtained will vary with the mean even though the covariance is held constant since the mean is modeled as a function of covariates. This assumption of equal covariance may be unrealistic and it may be necessary to have alternatives that relax this assumption.

Loss of efficiency is always an interesting phenomenon to investigate whenever pseudo-likelihood or estimating equations are used in place of the full likelihood approach. In some contexts, like binary data, for example, it has been found to vary from acceptable to ignorable (see, e.g., Geys *et al.*, 1998). In the context of count data however, there is need to quantify the efficiency lost and to understand the implications of our approaches.

Our proposals in Chapters 4 and 5 have not dealt with the very important and common aspect of longitudinal studies, namely, missing data. While we have assumed that the missingness mechanism has no relationship with the response variable of interest, this assumption may be overly restrictive and one which should be further investigated. It is necessary therefore to investigate our proposals for missingness implications along the lines of, for example, Molenberghs *et al.* (2011).

The data generator presented in Chapters 6 and 9 includes overdispersion as a time-independent phenomenon. Should one be interested in generating data with time-dependent overdispersion, extensions are necessary to take that into account. See Ye *et al.* (2013) for a case of modeling time-dependent overdispersion in longitudinal count data. Also, since the combined model is hierarchical in formulation, only positive correlated variables can be generated. A potential solution to the positive correlations restriction may be the generation directly from a multivariate model, in this case, the multivariate negative-binomial model (see, e.g., Solis-Trapala and Farewell, 2005).

The method of Ghidey *et al.* (2004) implemented in Chapter 10 can get computationally intensive depending on the size of the data set and the specified mean structure. Incorporating the computationally intensive part of the method in C++ may improve computation time.

Bibliography

Aerts M, Geys H, Molenberghs G, Ryan L (2002). *Topics in Modelling of Clustered Data*. Chapman & Hall/CRC, Boca Raton, FL.

Agresti A (2002). *Categorical Data Analysis*. 2nd edition. John Wiley & Sons, New York.

Arnold B, Strauss D (1991). "Pseudolikelihood Estimation: Some Examples." *Sankhya: the Indian Journal of Statistics - series B*, **53**, 233–243.

Avramidis A, Channouf N, L'Ecuyer P (2009). "Efficient Correlation Matching for Fitting Discrete Multivariate Distributions with Arbitrary Marginals and Normal-Copula Dependence." *INFORMS Journal on Computing*, **2**, 88–106.

Booth J, Casella G, Friedl H, Hobert J (2003). "Negative Binomial Loglinear Mixed Models." *Statistical Modelling*, **3**, 179–181.

Breslow N (1984). "Extra-Poisson Variation in Log-linear Models." *Applied Statistics*, **33**, 38–44.

Breslow N, Clayton D (1993). "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association*, **88**, 9–25.

Butler S, Louis T (1992). "Random Effects Models with Nonparametric Priors." *Statistics in Medicine*, **11**, 1981–2000.

Cameron A, Johansson P (1997). "Count Data Regression using Series Expansions: with Applications." *Journal of Applied Econometrics*, **12**, 203–223.

Cameron A, Trivedi P (2013). *Regression Analysis of Count Data*. 2nd edition. Cambridge University Press, Cambridge.

Cario M, Nelson B (1997). "Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix." *Technical report*, Northwestern University, Evanston, Illinois.

Cario M, Nelson B (1998). "Numerical Methods for Fitting and Simulating Autoregressive-to-anything Processes." *INFORMS Journal on Computing*, **10**, 72–81.

Castillo J, Pérez-Casany M (1998). "Weighted Poisson Distributions for Overdispersion and Underdispersion Situations." *Annals of the Institute of Statistical Mathematics*, **50**, 567–585.

Chernick M (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*. 2nd edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

Chin H, Quddus M (2003). "Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections." *Accident Analysis & Prevention*, **35**, 253–259.

Cressie N (1991). *Statistics for Spatial Data*. John Wiley & Sons, Inc., New York.

Crowder M (1995). "On the Use of a Working Correlation Matrix in Using Generalised Linear Models for Repeated Measures." *Biometrika*, **82**, 407–410.

Dale J (1986). "Global Cross-ratio Models for Bivariate, Discrete, Ordered Responses." *Biometrics*, **42**, 721–727.

Deb P, Holmes A (2000). "Estimates of Use and Costs of Behavioural Health Care: A Comparison of Standard and Finite Mixture Models." *Health Economics*, **9**, 475–489.

Devroye L (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

Diggle P, Heagerty P, Liang KY, Zeger S (2002). *Analysis of Longitudinal Data*. Oxford Science Publications, 2nd edition. Clarendon Press, Oxford.

Downer R, Moser E (2001). "On the Generation of a Multivariate Spatial Poisson Distribution." *Technical report*, Lousiana State University.

Eilers P, Marx B (1996). "Flexible Smoothing with B-Splines and Penalties (With Discussions)." *Statistical Science*, **11**, 89–121.

Fahrmeir L, Tutz G (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, Heidelberg.

Fahrmeir L, Tutz G (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer Series in Statistics, 2nd edition. Springer-Verlag, New York.

Fitzmaurice G, Laird N, Ware J (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, New York.

Geys H, Molenberghs G, Lipsitz S (1998). "A Note on the Comparison of Pseudo-likelihood and Generalized Estimating Equations for Marginal Odds Ratio Models." *Journal of Statistical Computation and Simulation*, **62**, 4572.

Geys H, Molenberghs G, Ryan L (1999). "Pseudolikelihood Modeling of Multivariate Outcomes in Developmental Toxicology." *Journal of the American Statistical Association*, **94**, 734–745.

Ghidey W (2005). *Relaxing the Normality Assumption of the Random Effects Distribution in the Linear Mixed Model*. Ph.D. thesis, Biostatistical Center, Katholieke Universiteit Leuven.

Ghidey W, Lesaffre E, Eilers P (2004). "Smooth Random Effects Distribution in a Linear Mixed Model." *Biometrics*, **60**, 945–953.

Ghosh S, Pasupathy R (2012). "C-NORTA: A Rejection Procedure for Sampling from the Tail of Bivariate NORTA Distributions." *INFORMS Journal on Computing*, **24**, 295–310.

Gray R (1992). "Flexible Methods for Analyzing Survival Data using Splines with Application to Breast Cancer Prognosis." *Journal of the American Statistical Association*, **87**, 942–951.

Hall D (2000). "Zero-inflated Poisson and Binomial Regression with Random Effects: A Case Study." *Biometrics*, **56**, 1030–1039.

Hall D, Zhang Z (2004). "Marginal Models for Zero Inflated Clustered Data." *Statistical Modelling*, **4**, 161–180.

Hardin J, Hilbe J (2003). *Generalized Estimating Equations*. Chapman and Hall/CRC, Boca Raton.

Harville DA (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." *Journal of the American Statistical Association*, **72**, 320–340.

Heagerty P (1999). "Marginally Specified Logistic-Normal Models for Longitudinal Binary Data." *Biometrics*, **55**, 688–698.

Hinde J, Demétrio C (1998a). "Overdispersion: Models and Estimation." *Computational Statistics and Data Analysis*, **27**, 151–170.

Hinde J, Demétrio C (1998b). "Overdispersion: Models and Estimation." *Technical report*, São Paulo.

Hofert M, Kojadinovic I, Maechler M, Yan J (2012). **Copula***: Multivariate Dependence with Copulas*. R Foundation for Statistical Computing. R package version 0.999-4, URL `http://CRAN.R-project.org/package=copula`.

Hofert M, Maechler M (2011). "Nested Archimedean Copulas Meet R: The **nacopula** Package." *Journal of Statistical Software*, **39(9)**, 1–20. URL `http://www.jstatsoft.org/v39/i09/`.

Iddi S, Molenberghs G (2013). "A Marginalized Model for Zero-inflated, Overdispersed and Correlated Count Data." *Electronic Journal of Applied Statistical Analysis*, **6**, 149–165.

Joe H, Lee Y (2008). "On Weighting of Bivariate Margins in Pairwise Likelihood." *Journal of Multivariate Analysis*, **100**, 670–685.

Jørgensen B (1987). "Exponential Dispersion Models." *Journal of the Royal Statistical Society, Series B*, **49**, 127–162.

Karlis D (2003). "An EM Algorithm for Multivariate Poisson Distribution and Related Models." *Journal of Applied Statistics*, **30**, 63–77.

Karlis D, Ntzoufras I (2003). "Analysis of Sports Data by using Bivariate Poisson Models." *The Statistician*, **52**, 381–393.

Kassahun W, Neyens T, Molenberghs G, Faes C, Verbeke G (2012). "Modeling Overdispersed Longitudinal Binary Data using a Combined Beta and Normal Random-effects Model." *Archives of Public Health*, **70**. Issue: 1, Chapter: 7, Article: 7.

Kim H, Shults J (2010). "%QLS SAS Macro: A SAS Macro for Analysis of Correlated Data Using Quasi-Least Squares." *Journal of Statistical Software*, **35**, 1–22.

Kocherlakota S, Kocherlakota K (1992). *Bivariate Discrete Distributions*. CRC Press, Boca Raton.

Kocherlakota S, Kocherlakota K (2001). "Regression in the Bivariate Poisson Distribution." *Communications in Statistics, Theory & Methods*, **30**, 815–825.

Kojadinovic I, Yan J (2010). "Modeling Multivariate Distributions with Continuous Margins using the **copula** R Package." *Journal of Statistical Software*, **34(9)**, 1–20. URL `http://www.jstatsoft.org/v34/i09/`.

Laird N, Ware J (1982). "Random-effects Models for Longitudinal Data." *Biometrics*, **38**, 963–974.

Lakshminarayana J, Pandit S, Rao K (1999). "On a Bivariate Poisson Distribution." *Communications in Statistics, Theory & Methods*, **28**, 267–276.

Lambert D (1992). "Zero-Infated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics*, **34**, 1–13.

Lawless J (1987). "Negative Binomial and Mixed Poisson Regression." *The Canadian Journal of Statistics*, **15**, 209–225.

Le Cessie S, Van Houwelingen J (1994). "Logistic Regression for Correlated Binary Data." *Applied Statistics*, **43**, 95–108.

Lee Y, Nelder J (2004). "Conditional and Marginal Models: Another View." *Statistical Science*, **19**, 219–228.

Leppik IE, Dreifuss F, Bowman-Cloyd T, et al (1985). "A Double-blind Crossover Evaluation of Progabide in Partial Seizures." *Neurology*, **35**, 285.

Lesaffre E, Asefa M, Verbeke G (1999). "Assessing the Goodness-of-fit of the Laird and Ware Model-An Example: the Jimma Infant Survival Differential Longitudinal Study." *Statistics in Medicine*, **18**, 835–854.

Li S, Hammond J (1975). "Generation of Pseudo-random Numbers with Specified Univariate Distributions and Correlation Coefficients." *IEEE Transactions on Systems, Man, and Cybernetics*, **5**, 557–561.

Liang K, Zeger S (1986). "Longitudinal Data Analysis using Generalized Linear Models." *Biometrika*, **73**, 13–22.

Liang K, Zeger S, Qaqish B (1992). "Multivariate Regression Analyses for Categorical Data." *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

Lindsey JK, Lambert P (1998). "On the Appropriateness of Marginal Models for Repeated Measurements in Clinical Trials." *Statistics in Medicine*, **17**, 447–469.

Lipsitz S, Laird N, Harrington D (1991). "Generalized Estimating Equations for Correlated Binary Data: using the Odds Ratio as a Measure of Association." *Biometrika*, **78**, 153–160.

Madsen L, Dalthorp D (2007). "Simulating Correlated Count Data." *Environmental and Ecological Statistics*, **14**, 129–148.

Mardia K (1970). *Families of Bivariate Distributions*. Griffin, London.

McCullagh P, Nelder J (1989). *Generalized Linear Models*. Chapman & Hall, London.

McCulloch C, Searle S (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.

Molenberghs G, Kenward M (2010). "Semi-parametric Marginal Models for Hierarchical Data and their Corresponding Full Models." *Computational Statistics & Data Analysis*, **54**, 585–597.

Molenberghs G, Kenward M, Verbeke G, Teshome Ayele B (2011). "Pseudo-likelihood Estimation for Incomplete Data." *Statistica Sinica*, **21**, 187–206.

Molenberghs G, Lesaffre E (1994). "Marginal Modeling of Correlated Ordinal Data using a Multivariate Plackett Distribution." *Journal of the American Statistical Association*, **89**, 633644.

Molenberghs G, Verbeke G (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, New York.

Molenberghs G, Verbeke G (2011). "A Note on the Hierarchical Interpretation for Negative Variance Components." *Statistical Modeling*, **11**, 389–408.

Molenberghs G, Verbeke G, Demétrio C (2007). "An Extended Random Effects Approach to Modeling Repeated Overdispersed Count Data." *Lifetime Data Analysis*, **13**, 513–531.

Molenberghs G, Verbeke G, Demétrio C, Vieira A (2010). "A Family of Generalized Linear Models for Repeated Measures with Normal and Conjugate Random Effects." *Statistical Science*, **25**, 325–347.

Nelder J, Wedderburn R (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society B*, **135**, 370–384.

Nelsen R (2006). *An Introduction to Copulas*. Springer-Verlag, Berlin.

Park C, Shin D (1998). "An Algorithm for Generating Correlated Random Variables in a Class of Infinitely Divisible Distributions." *Journal of Statistical Computation and Simulation*, **61**, 127–139.

Parzen M, Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Troxel A, Molenberghs G (2007). "Pseudo-likelihood Methods for the Analysis of Longitudinal Binary Data Subject to Nonignorable Non-monotone Missingness." *Journal of Data Science*, **5**, 1–21.

Powers D, Xie Y (2008). *Statistical Methods for Categorical Data Analysis*. 2nd edition. Emerald Group publishing Limited, Howard House, UK.

Prentice R (1988). "Correlated Binary Regression with Covariates Specific to each Binary Observation." *Biometrics*, **44**, 1033–1048.

Prentice R, Zhao L (1991). "Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Response." *Biometrics*, **47**, 825–839.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

SAS Institute Inc (2011). *SAS/IML9.3 User's Guide*. Cary, NC. URL `http://www.sas.com/`.

Pryseley A, Tchonlafi C, Verbeke G, Molenberghs G (2011). "Estimating Negative Varinace Components from Gaussian and Non-Gaussian Data: A Mixed Models Approach." *Computational Statistics and Data Analysis*, **55**, 1071–1085.

Ridout M, Besbeas P (2004). "An Empirical Model for Underdispersed Count Data." *Statistical Modelling*, **4**, 77–89.

Ridout M, Hinde J, Demetrio C (2001). "A Score Test for a Zero-inflated Poisson Regression Model against Zero-inflated Negative Binomial Alternatives." *Biometrics*, **57**, 219–233.

SAS Institute Inc (2002-2004). *SAS 9.1.3 Help and Documentation*. Cary, NC. URL `http://www.sas.com/`.

Sellers K, Shmueli G (2010). "A Flexible Regression Model for Count Data." *The Annals of Applied Statistics*, **4**, 943–961.

Shin K, Pasupathy R (2010). "An Algorithm for Fast Generation of Bivariate Poisson Random Vectors." *INFORMS Journal on Computing*, **22**, 81–92.

Skellam J (1948). "A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials." *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.

Solis-Trapala I, Farewell V (2005). "Regression Analysis of Overdispersed Correlated Count Data with Subject Specific Covariates." *Statistics in Medicine*, **24**, 25572575.

Sun W, Shults J, Leonard M (2009). "A Note on the Use of Unbiased Estimating Equations to Estimate Correlation in Analysis of Longitudinal Trials." *Biometrical Journal*, **51**, 5–18.

Sutradar B (2011). *Dynamic Mixed Models for Familial Longitudinal Data*. Springer-Verlag, New York.

Thall P, Vail S (1990). "Some Covariance Models for Longitudinal Count Data with Overdispersion." *Biometrics*, **46**, 657–671.

Troxel A, Lipsitz S, Harrington D (1998). "Marginal Models for the Analysis of Longitudinal Measurements with Nonignorable Non-monotone Missing Data." *Biometrika*, **85**, 661–672.

van Iersel M, Oetting R, Hall DB (2000). "Imidacloprid Applications by Subirrigation for Control of Silverleaf Whitefly (Homoptera: Aleyrodidae) on Poinsettia." *Journal of Economic Entomology*, **93**, 813–819.

Verbeke G, Lesaffre E (1996). "A Linear Mixed Model with Heterogeneity in the Random-effects Population." *Journal of the American Statistical Association*, **91**, 217–221.

Verbeke G, Lesaffre E (1997). "The Effect of Misspecifying the Random Effects Distribution in the Linear Mixed Models for Longitudinal Data." *Computational Statistics and Data Analysis*, **23**, 541–556.

Verbeke G, Molenberghs G (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

Wang YG, Carey VJ (2004). "Unbiased Estimating Equations from Working Correlation Models for Irregularly Timed Repeated Measures." *Journal of the American Statistical Association*, **99**, 845–852.

Wicklin R (2013). "What versions of R are supported by SAS?" URL `http://blogs.sas.com/content/iml/2013/09/16/what-versions-of-r-are-supported-by-sas/`.

Winkelmann R (2004). "Health Care Reform and the Number of Doctor Visits. An Econometric Analysis." *Journal of Applied Econometrics*, **19**, 455–472.

Winkelmann R (2008). *Econometric Analysis of Count Data*. Springer, Berlin.

Wolfinger R, O'Connell M (1993). "Generalized Linear Mixed Models: A Pseudo-likelihood Approach." *Journal of Statistical Computing and Simulation*, **48**, 233–243.

Yahav I, Shmueli G (2012). "On Generating Multivariate Poisson Data in Management Science Applications." *Applied Stochastic Models for Business and Industry*, **28**, 91–102.

Yan J (2007). "Enjoy the Joy of Copulas: With a Package **copula**." *Journal of Statistical Software*, **21**, 1–21. URL `http://www.jstatsoft.org/v21/i04/`.

Ye F, Yue C, Yang Y (2013). "Modeling Time-dependent Overdispersion in Longitudinal Count Data." *Computational Statistics and Data Analysis*, **58**, 257–264. URL `http://dx.doi.org/10.1016/j.csda.2012.08.009`.

Yi G, Zeng L, Cook R (2011). "A Robust Pairwise Likelihood Method for Incomplete Longitudinal Data Arising in Clusters." *Canadian Journal of Statistics*, **39**, 34–51.

Zeileis A, Kleiber C, Jackman S (2008). "Regression Models for Count Data in R." *Journal of Statistical Software*, **27(8)**.

Zhao L, Prentice R (1990). "Correlated Binary Regression using a Quadratic Exponential Model." *Biometrika*, **77**, 642–648.

Zhao Y, Joe H (2005). "Composite Likelihood Estimation in Multivariate Data Analysis." *Canadian Journal of Statistics*, **33**, 335–356.

Ziegler A (2011). *Generalized Estimating Equations*. Springer, New York.

Part V

Supplementary Material

Appendix A

Supplementary Material for Chapter 4

A.1 Consistency and Asymptotic Normality of the Pseudo-likelihood Estimator

We first list the required regularity conditions on the density functions $f_s(\boldsymbol{y}^{(s)}; \lambda)$.

A0 The densities $f_s(\boldsymbol{y}^{(s)}; \lambda)$ are distinct for different values of the parameter λ.

A1 The densities $f_s(\boldsymbol{y}^{(s)}; \lambda)$ have common support, which does not depend on λ.

A2 The parameter space Ω contains an open region ω of which the true parameter value λ_0 is an interior point.

A3 ω is such that for all s, and almost all $\boldsymbol{y}^{(s)}$ in the support of $Y^{(s)}$, the densities admit all third derivatives

$$\frac{\partial^3 f_s(\boldsymbol{y}^{(s)}; \lambda)}{\partial\theta_j \partial\theta_k \partial\theta_\ell}.$$

A4 The first and second logarithmic derivatives of f_s satisfy

$$E_\lambda\left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)};\lambda)}{\partial\theta_k}\right)=0, \qquad k=1,\ldots,q,$$

and

$$0<E_\lambda\left(\frac{-\partial^2 \ln f_s(\boldsymbol{y}^{(s)};\lambda)}{\partial\theta_k\partial\theta_\ell}\right)<\infty, \qquad k,\ell=1,\ldots,q.$$

A5 The matrix I_0, defined in (A.1), is positive definite.

A6 There exist functions M_{klr} such that

$$\sum_{s\in S}\delta_s E_\lambda\left|\frac{\partial^3 \ln f_s(\boldsymbol{y}^{(s)};\lambda)}{\partial\theta_k\partial\theta_\ell\partial\theta_r}\right|<M_{k\ell r}(\boldsymbol{y})$$

for all $\boldsymbol{y}$ in the support of f and for all $\theta\in\omega$ and $m_{k\ell r}=E_{\lambda_0}(M_{k\ell r}(Y))<\infty$.

Theorem 1, proven by Arnold and Strauss (1991), guarantees the existence of at least one solution to the pseudo-likelihood equations, which is a consistent and asymptotically normal estimator. Without loss of generality, we can assume λ is constant. Replacing it by λ_i, and modeling it as a function of covariates is straightforward.

Theorem 1 (Consistency and Asymptotic Normality) *Assume that $(Y_1,\ldots,Y_N)$ are i.i.d. with common density that depends on λ_0. Then under regularity conditions (A1)–(A6):*

1. *the pseudo-likelihood estimator $\tilde{\lambda}_N$, defined as the maximizer of the pseudo-score function, converges in probability to λ_0.*
2. *$\sqrt{N}(\tilde{\lambda}_N-\lambda_0)$ converges in distribution to $N_p(\mathbf{0}, I_0(\lambda_0)^{-1}I_1(\lambda_0)I_0(\lambda_0)^{-1})$ with $I_0(\lambda)$ defined by*

$$I_{0,k\ell}(\lambda)=-\sum_{s\in S}\delta_s E_\lambda\left(\frac{\partial^2 \ln f_s(\boldsymbol{y}^{(s)};\lambda)}{\partial\theta_k\partial\theta_\ell}\right) \tag{A.1}$$

and $I_1(\lambda)$ *by*

$$I_{1,k\ell}(\lambda) = \sum_{s,t\in S} \delta_s \delta_t E_\lambda \left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)};\lambda)}{\partial \theta_k} \frac{\partial \ln f_t(\boldsymbol{y}^{(t)};\lambda)}{\partial \theta_\ell} \right). \tag{A.2}$$

A.2 The First and Second Derivatives of the Log Pseudo-likelihood Function

Let

$$\mathbf{B} = \sum_{k=0}^{min(y_{is},y_{it})} \frac{e^{\boldsymbol{X}_{is}\boldsymbol{\beta}(y_{is}-k)+\boldsymbol{X}_{it}\boldsymbol{\beta}(y_{it}-k)}\theta_{ist}^k}{(y_{is}-k)!(y_{it}-k)!k!}.$$

Then, the bivariate Poisson distribution for the two measurements y_{is} and y_{it} expressed in terms of the covariates at the two time points s and t is

$$f(y_{is}, y_{it}) = \exp\left[-\left(e^{\boldsymbol{X}_{is}\boldsymbol{\beta}} + e^{\boldsymbol{X}_{it}\boldsymbol{\beta}} + \theta_{ist}\right)\right] \times \mathbf{B}. \tag{A.3}$$

This leads to the log PL function given as

$$p\ell(\boldsymbol{\lambda}|\mathbf{Y}) = \sum_{i=1}^{K} \sum_{s<t} \log f(y_{is}, y_{it})$$

from which the gradient and Hessian functions are derived with respect to $\boldsymbol{\beta}$ and θ_{ist} (θ_{st} here) as

$$\begin{aligned} \frac{\partial p\ell}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{K} \sum_{s<t} \left\{ -\left(\boldsymbol{X}_{is}^T e^{\boldsymbol{X}_{is}\boldsymbol{\beta}} + \boldsymbol{X}_{it}^T e^{\boldsymbol{X}_{it}\boldsymbol{\beta}}\right) + \mathbf{B}^{-1} A \right\} \\ \frac{\partial p\ell}{\partial \theta_{st}} &= \sum_{i=1}^{K} \sum_{s<t} \left\{ -1 + \mathbf{B}^{-1}\mathbf{C}_2 \right\} \end{aligned} \tag{A.4}$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}}\Big(\frac{\partial p\ell}{\partial \boldsymbol{\beta}}\Big) = \sum_{i=1}^{K}\sum_{s<t}\left\{\begin{array}{l} -(\boldsymbol{X}_{is}^T\boldsymbol{X}_{is}e^{\boldsymbol{X}_{is}\boldsymbol{\beta}} + \boldsymbol{X}_{it}^T\boldsymbol{X}_{it}e^{\boldsymbol{X}_{it}\boldsymbol{\beta}}) + \\ \mathbf{B}^{-2}(A_d\mathbf{B} - \mathbf{A}\mathbf{A}^T) \end{array}\right\}$$

$$\frac{\partial}{\partial \theta_{st}}\Big(\frac{\partial p\ell}{\partial \theta_{st}}\Big) = \sum_{i=1}^{K}\sum_{s<t}\mathbf{B}^{-2}(\mathbf{B}\mathbf{C}_3 - \mathbf{C}_2^2) \tag{A.5}$$

$$\frac{\partial}{\partial \theta_{st}}\Big(\frac{\partial p\ell}{\partial \boldsymbol{\beta}}\Big) = \sum_{i=1}^{K}\sum_{s<t}\mathbf{B}^{-2}(\mathbf{B}\mathbf{C} - \mathbf{C}_2\mathbf{A})$$

where

$$\begin{aligned}
\mathbf{A}_1 =& e^{\boldsymbol{X}_{is}\boldsymbol{\beta}(y_{is}-k)+\boldsymbol{X}_{it}\boldsymbol{\beta}(y_{it}-k)}\\
\mathbf{A}_2 =& (y_{is}-k)\boldsymbol{X}_{is}^T + (y_{it}-k)\boldsymbol{X}_{it}^T\\
\mathbf{A} =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{\theta_{st}^k}{(y_{is}-k)!(y_{it}-k)!k!}\mathbf{A}_1\mathbf{A}_2\\
\mathbf{A}_d =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{\theta_{st}^k}{(y_{is}-k)!(y_{it}-k)!k!}\mathbf{A}_2\mathbf{A}_2^T\mathbf{A_1}\\
\mathbf{C} =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{k\theta_{st}^{k-1}\mathbf{A}_1\mathbf{A}_2}{(y_{is}-k)!(y_{it}-k)!k!}\\
\mathbf{C}_2 =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{k\theta_{st}^{k-1}\mathbf{A}_1}{(y_{is}-k)!(y_{it}-k)!k!}\\
\mathbf{C}_3 =& \sum_{k=0}^{min(y_{is},y_{it})} \frac{\mathbf{A}_1}{(y_{is}-k)!(y_{it}-k)!k!}k(k-1)\theta_{st}^{k-2}
\end{aligned}$$

A.3 Covariance Parameter Constrained to be Positive

We present an additional Table A.1 related to the hierarchical interpretation, where θ_{st} is constrained to be strictly positive.

Table A.1: *Simulation study, association: Parameter estimates, MSE and convergence rate of pseudo-likelihood for varying number of measurements per subject (n_i) and sample size (K), when the covariance(θ_{st}) is constrained to be positive*

		Parameter Estimates					MSE				
K	n_i	β_0	β_1	β_2	β_3	θ_{st}	β_0	β_1	β_2	β_3	$RATE_c$
True value		1.5807	-0.1881	-0.0340	0.0192	.					
10	2	1.0275	-0.3366	-0.1742	0.1689	1.9424	12.7503	18.6687	5.6976	8.0308	93
10	4	1.2937	-0.2635	-0.0534	0.0362	1.5647	0.3822	0.5165	0.0271	0.0497	98
10	8	1.3449	-0.2225	-0.0473	0.0241	1.3904	0.2210	0.3112	0.0033	0.0052	100
10	16	1.4176	-0.2365	-0.0473	0.0256	1.2149	0.1379	0.2314	0.0006	0.0007	100
100	2	1.2493	-0.2745	-0.0561	0.0318	1.9167	0.1760	0.1524	0.0236	0.0506	100
100	4	1.2811	-0.2670	-0.0528	0.0282	1.8182	0.1126	0.0592	0.0023	0.0046	100
100	8	1.3175	-0.2615	-0.0519	0.0283	1.7079	0.0821	0.0364	0.0006	0.0006	100
100	16	1.3923	-0.2709	-0.0509	0.0284	1.4696	0.0434	0.0281	0.0003	0.0002	100
1,000	2	1.2591	-0.2709	-0.0535	0.0293	1.9175	0.1094	0.0204	0.0022	0.0046	100
1,000	4	1.2807	-0.2691	-0.0533	0.0284	1.8555	0.0924	0.0119	0.0006	0.0005	100
1,000	8	1.3179	-0.2713	-0.0524	0.0290	1.7340	0.0703	0.0099	0.0004	0.0002	100
1,000	16	1.3907	-0.2737	-0.0512	0.0287	1.4994	0.0369	0.0095	0.0003	0.0001	100
10,000	2	1.2626	-0.2709	-0.0547	0.0296	1.9194	0.1018	0.0083	0.0006	0.0006	100
10,000	4	1.2811	-0.2709	-0.0534	0.0290	1.8570	0.0900	0.0074	0.0004	0.0001	100

Summary - Samenvatting

Summary

Many methods in general have been proposed for the analysis or generation of hierarchical data. Specific to count data, such methods include the generalized estimating equations, zero-inflated Poisson, zero-inflated negative-binomial models, the generalized linear mixed model and the more general combined model. These tools span a variety of modeling frameworks, namely, marginal, conditionally specified and subject-specific models. When interested in making inference on a (sub)population, marginal models are the way to go. GEE has been received in the world of research and it works fine when scientific interest is only in making inference about the marginal mean parameters. Should the researcher aim at making inference also on the association structure, GEE falls short because it allows the misspecification on the working correlation structure implying that that structure is a nuisance.

In Chapter 4, we have endeavored to propose a viable alternative to GEE through which inference can be made on the mean as well as the association. This was done using pseudo-likelihood mechanisms in which a pairwise likelihood was constructed based on the bivariate Poisson distribution and the parameters of interest obtained by maximizing the pseudo log-likelihood. Chapter 5 yet again proposed another alternative to GEE that permits inference on both the marginal mean parameters and the association. This involved the solving of score equations constructed at the level of each pair of observations from a subject while the standard errors were calculated using a sandwich estimator similar to that used in Chapter 4 or GEE. The bivariate Poisson distribution was used here as well which implies that the covariance parameter in the bivariate Poisson distribution would quantify the relationship between two observations of a subject.

The need to evaluate the performance of the proposals in Chapters 4 and 5 motivated an investigation into methods for simulating correlated count data. The combined model also turned out to be useful not only for the simultaneous modeling of correlated and overdispersed data but also in the context of data generation. With a pre-specification of the desired marginal mean (in terms of covariates) and covariance structure, correlated count data can be generated

from the combined model.

It was part of the agenda of this thesis to provide the software that can be used to implement our proposals. SAS macros have been therefore written with the goal of making it easy for the user to run or apply our proposals. Special attention was given to the simplification and automation of the creation of dummy variables in the presence of classification variables using existing SAS procedures. Therefore, Chapters 7, 8 and 9 describe SAS macros corresponding to the proposals in Chapters 4, 5 and 6, respectively.

Finally, we have also provided a SAS macro implementing the method of Ghidey *et al.* (2004) in the context of the linear mixed model. This method fits a linear mixed model but with a more flexible and general distribution function for the random effects, entirely based on the assumption that the density of the random effects can be well approximated by a mixture of Gaussian densities defined on a fine grid. Thus, this thesis has contributed to the literature of marginal models in the context of correlated count data by proposing a pseudo-likelihood approach and an extension of the second-order generalized estimating equations to permit inference on both the marginal mean and association parameters. It has also presented the combined model as an alternative tool via which correlated counts can be generated and has provided the corresponding SAS software implementing these approaches. In addition, a SAS macro has been presented that flexibly fits the linear mixed model while relaxing the usually made assumption of normality for the random effects.

Samenvatting

Nogal wat methoden werden voorgesteld in de literatuur voor enerzijds analyse en anderzijds generatie van hiërarchische gegevens. In het bijzonder voor aantallen omvatten dergelijke methoden: veralgemeende schattingsvergelijkingen (generalized estimating equations, GEE), Poisson modellen met zero-inflatie, het veralgemeend lineaire gemengd model, en het algemenere gecombineerde model. Deze methoden omvatten nemen een plaats in binnen verschillende modelmatige kaders, d.w.z., marginale, conditionele en subject-

specifieke modellen. Wanneer de wetenschappelijke vraag zich richt op een (deel)populatie eerder dan een individueel subject, dan zijn marginale modellen meest aangewezen. GEE is hiertoe reeds lang goed onthaald en veelgebruikt in de onderzoeksgemeenschap; de methode werkt meer dan behoorlijk wanneer inferentie zich beperkt tot parameters in het marginale gemiddelde. Wanneer de onderzoeker eveneens wetenschappelijke vragen heeft omtrent de associatie-structuur, dan is GEE niet geschikt omdat het foutief gespecifieerde correlatiestructuren toelaat. In GEE is correlatie dus een zogenaamd *nuisance* kenmerk.

In Hoofdstuk 4 hebben we een werkbaar alternatief voorgesteld voor GEE, waardoor inferentie mogelijk is voor zowel gemiddelde als associatie. Hiervoor gebruikten we pseudo-likelihood; meerbepaald werd een paarsgewijze likelihood geconstrueerd op basis van de bivariate Poisson verdeling. De belangrijke parameters werden geschat door het maximaliseren van de pseudo log-likelihood. Daarnaast stelde Hoofdstuk 5 een bijkomend GEE alternatief voor dat inferentie toelaat in zowel het marginale gemiddelde als de associatieparameters. Hiertoe worden scorevergelijkingen opgelost, op het niveau van elk observatiepaar; standaardfouten volgen uit het gebruik van de sandwichschatter, naar analogie met deze gebruikt in Hoofdstuk 4 en in gewone GEE. De bivariate Poisson verdeling werd ook hier gebruikt. Dit impliceert dat de covariantieparameter in deze verdeling kan gebruikt worden om de relatie tussen twee observaties, opgetekend aan hetzelfde subject te kwantificeren.

De nood aan performantie-evaluatie van de voorstellen uit Hoofdstukken 4 en 5 heeft ons gemotiveerd om simulatiestudies te ondernemen. In die zin was het gecombineerde model nuttig, niet alleen om gecorreleerde gegevens met overdispersie te modelleren, maar ook om gegevens te genereren. Vertrekkende van de specificatie van het gewenste marginale gemiddelde en van de variantie-covariantie structuur, waarbij het gemiddelde van covariaten kan afhangen, kan men nu makkelijk aantallen simuleren op basis van het gecombineerde model.

Bij de voorgaande ontwikkelingen was het ook expliciet voorzien van de nodige gebruiksvriendelijke software te ontwikkelen. SAS macro's wer-

den vanuit dit oogmerk ontwikkeld. Hierbij waren flexibiliteit zowel als gebruiksvriendelijkheid uitgangspunten. Er werd aandacht besteed aan het creëren van dummy veranderlijken wanneer men gebruikt maakt van classificatie veranderlijken, ook met bestaande SAS procedures. Vanuit dit standpunt stellen Hoofdstukken 7, 8 en 9 SAS macro's voor, die respectievelijk overeenkomen met de methodologie in Hoofdstukken 4, 5, en 6.

Tot slot hebben we een SAS macro ontwikkeld die de methode van Ghidey *et al.* (2004) implementeert in de context van het lineair gemengde model. Dit instrument fit een lineair gemengd model, maar met een flexibelere verdeling voor de random effecten. Er wordt vertrokken vanuit het standpunt dat de dichtheid van de random effecten goed kan benadered worden door een mengeling van normale dichtheden, over een fijn rooster.

We durven daarom besluiten dat dit proefschrift bijgedragen heeft aan de literatuur omtrent marginale modellen voor herhaald gemeten aantallen, door gebruik te maken van pseudo-likleihod en door uitbreiding van gewone naar tweede-orde GEE. Hierdoor kunnen we conclusies trekken, niet alleen voor de marginale gemiddelde functies, maar evenzeer voor associatieparameters. Dit werk stelde eveneens het gecombineerde model voor als een instrument om herhaalde aantallen te genereren, waar bovendien wordt rekening gehouden met overdispersie. Voor alle bestudeerde en ontwikkelde methodologie werden SAS macro's ontwikkeld.

ACTA BIOMEDICA LOVANIENSIA

597. H. VAN REMOORTEL, Physical Activity and Comorbidities in Patients with COPD. 2013
598. G. J. UGHI, Automated Analysis of Intracoronary Optical Coherence Tomography Images. 2013
599. K. STAATS, Mechanisms in Amyotrophic Lateral Sclerosis: From Genetic Linkage to Novel Insights. 2013
600. A. G. MONEA, Biomechanical Characterization of Bicycle Accidents Related Head Injuries. 2013
601. G. VAN DER MIEREN, The Effect of Type II Diabetes and the Metabolic Syndrome on Cardiac Second Window of Preconditioning. 2013
602. P. GOEMINNE, (Non)-Cystic Fibrosis Bronchiectasis. Mortality, Makers and Markers of Disease. 2013
603. I. CALLEBAUT, Naso-Ocular Interaction In Allergic Rhinitis. 2013
604. H. MAURIN, Inter-Relations of Protein Tau, GSK3 and Nectin-3 in Synaptic Structure and Plasticity in the Hippocampus of Mouse Models for Alzheimer's Disease. 2013
605. P. THEVISSEN, Dental Age Estimation in Sub-Adults: Striving for an Optimal Approach. 2013
606. K. DECKERS, Antigen-Specific Suppression of EAE Using Immunization with MHC Class II-Restricted Autoantigen Epitopes Containing an Oxido-Reductase Motif. 2013
607. C. LAMBRECHT, Characterization of Tumorigenesis in the PP2A B'δ Knockout Mouse: A Model for Hepatocarcinogenesis. 2013
608. I. MYATCHIN, Working Memory in Children with Epilepsy: an Event-Related Potentials Study. 2013
609. C. THEUNIS, Development and Validation of Active and Passive Immunotherapy Targeted at Protein Tau in Transgenic Mouse Models for Alzheimer's Disease. 2013
610. V. A. PICAVET, The Hidden Danger of BDD in Aesthetic Rhinoplasty: Prospective Outcome Studies in Aesthetic Nose Surgery with Emphasis on the Major Impact of Body Dysmorphic Disorder. 2013
611. L. CLINCKEMALIE, The TMPRSS2 Gene: Study of the Androgen Regulation and the Effects of Genetic Polymorphisms on Androgen Receptor Binding. 2013
612. K. JANSEN, Multimodal Monitoring of Electrophysiological Signals in Childhood Epilepsy and Neonatal Encephalopathy. 2013
613. C. TIEDTKE, Return to Work Experiences after Breast Cancer. 2013
614. Y. FENG, Development, Characterization and Applications of a Rabbit Model of Ischemic Heart Disease with Multiparametric Imaging Biomarkers. 2013
615. D. BARBOSA, Automated Assessment of Cardiac Morphology and Function. 2013
616. S. BOBIĆ, Role of VEGF Family in Inflammation and Edema of Upper and Lower Airways. 2013
617. M. MIRANDA CONA, Radiopharmaceutical Research on Selected Oncological and Non-Oncological Topics in Translational Medicine. 2013
618. Y-M. VANWIJNGAERDEN, Pathophysiology of Cholestatic Liver Dysfunction during Critical Illness. 2013
619. A.-S. PAPADOPOULOU, Physiological Roles of miR-29. 2013
620. S.A.M. SHARIATI, APLP2 Regulates Neuronal Stem Cell Differentiation during Cortical Development. 2013

621. S. GOETHALS, Nurses' Ethical Reasoning and Behaviour in Cases of Physical Restraint in Acute Elderly Care in Flanders: a Qualitative Empirical Study. 2013
622. V. FERFERIEVA, Ultrasonic Deformation Imaging in a Small Animal Model of Chronic Myocardial Infarction. 2013
623. M. BASTOS MARQUES, The Potentially Protective Role of Adipose Tissue during Critical Illness. 2013
624. I. VANHEES, Bone Loss in Critical Illness: From Molecular Pathways to *in vivo* Analysis. 2013
625. S. SEYS, Asthma: From Airway Inflammation to Phenotypes. 2013
626. L. TONG, Novel Beam Forming Methods for Fast Cardiac Imaging Using Ultrasound. 2013
627. R. ABDOLLAHI ASADABADI, Action Observation and Retinotopy, Functional Imaging of the Human Visual Cortex. 2013
628. I. VOGEL, Costimulatory Signals and Activation of Regulatory T Cells. 2013
629. T. ADRIAENSSENS, Assessment of Vessel Wall Healing after Percutaneous Coronary Intervention Using Optical Coherence Tomography. 2014
630. R. JASAITYTĖ, A Novel Echocardiographic Index for the Estimation of Left Ventricular Contractility. 2014
631. R. BRUFFAERTS, Distributed Processing of Concrete Entities in Temporal Neocortex. 2014
632. T. C. F. FONSECA, Improvement of *in vivo* Measurements by Development and Application of a 3D Human Body Library Based on Polygonal Mesh Surfaces for WBC Set-Up Calibration. 2014
633. O. APANASETS (IVASHCHENKO), Interplay between Peroxisome Biogenesis and Redox Balance in Mammalian Cells. 2014
634. S. VERELST, Emergency Department Crowding in Relation to In-Hospital Adverse Medical Events. 2014
635. F. RENZI, Pat1b and LSm14 Proteins: Regulators of Neuronal mRNA Metabolism with a Possible Role in Spinal Muscular Atrophy. 2014
636. C. BREYNAERT, Studies on Inflammation and Fibrosis in a Model of Chronic Inflammatory Colitis. 2014
637. J. QIAN, Mitotic Function and Regulation of the Phosphatase Scaffold Repo-Man. 2014
638. E. SHAHEEN, The Development of a Methodology to Optimize the Performance of Breast Tomosynthesis. 2014
639. A. PIMENOVA, Regulation of Amyloid Precursor Protein Processing by Serotonin Signaling. 2014
640. L. MAHIEU, Care for Older People with Regard to Intimacy and Sexuality: A Clinical-Ethical Study with Special Attention to Institutionalized People with Dementia. 2014
641. H. VANHEEL, Duodenal Implications in the Pathophysiology of Functional Dyspepsia: Focus on Impaired Mucosal Integrity and Low-Grade Inflammation. 2014
642. A. ALQERBAN, Maxillary Canine Impaction and Adjacent Incisor Root Resorption. 2014
643. L. DE MUYNCK, Progranulin and C9orf72 in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. 2014
644. T. BISELELE, New Approach of Diagnosis and Therapy of Perinatal Asphyxia in the Democratic Republic of Congo. 2014
645. E. BOONEN, Novel Insights in the HPA-axis during Critical Illness. 2014
646. R. SONNEVILLE, Glucose Neurotoxicity in Critical Illness. 2014
647. G. KALEMA, Flexible Regression Models for the Analysis of Hierarchical Data from Medical Studies. 2014